DH@Madrid Summer School 2021

Herramientas Digitales para las Humanidades Digitales en la e-infraestructura CLARIN

Mikel Iruskieta
Ixa Taldea - HiTZ zentroa
UPV/EHU

www.clarin.eu
www.clarin-es.org
http://ixa2.si.ehu.es/clarink

Creación de un Proyecto de Humanidades Digitales basado en el análisis de textos: Modelado y Procesamiento

UNED
Fundación Uned
LiNHD LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES
eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology

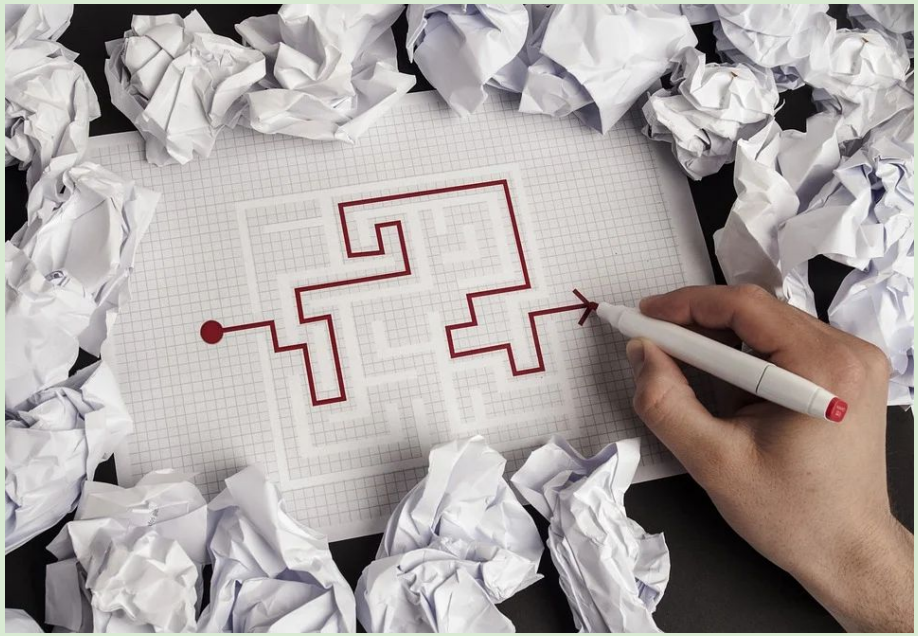
Esquema

1. Resumen de la sesión
2. Introducción: e-Infraestructuras de investigación
   a. Acciones conjuntas de infraestructuras
3. Justificación
4. Casos prácticos
   a. Herramientas de análisis textual
   b. Herramientas de análisis de voz
   c. Casos de uso prácticos
   d. Otros recursos
5. Conclusiones

Mapa

Resumen

- EOSC e Infraestructuras
- Interoperabilidad (dentro y fuera)
- Justificación de infraestructuras
- Casos de uso



Resumen de la sesión

Objetivo

- *European Open Science Cloud*: EOSC
- Ciencia de principios FAIR: Encontrable, Accesible, Interoperable y Reutilizable
- Interoperabilidad en infraestructuras (CLARIN)

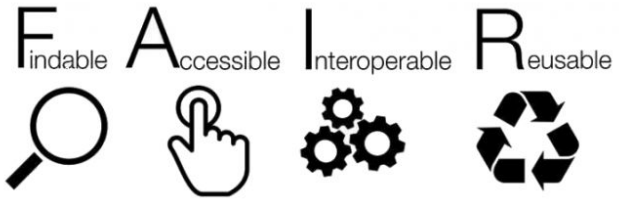
Método

  - Construir infraestructura que no se desarrollará en Europa para
    - Lenguas oficiales y cooficiales del estado
  - ALL-LT-in-ONE-URL: todos los recursos, todos los servicios

Ejemplos

- Casos de uso y herramientas (sencillas) para investigar en infraestructuras europeas que se inter-comunican

EUROPEAN OPEN SCIENCE CLOUD



Justificación de las infraestructuras

Ciencia: abierta de impacto y reproducible:

- Fragmentación en la investigación de CCSS
- Recursos de CCSS dispersados en repositorios
- Poco re-uso en la investigación de CCSS
- Poca interdisciplinaridad
- Impacto social limitado

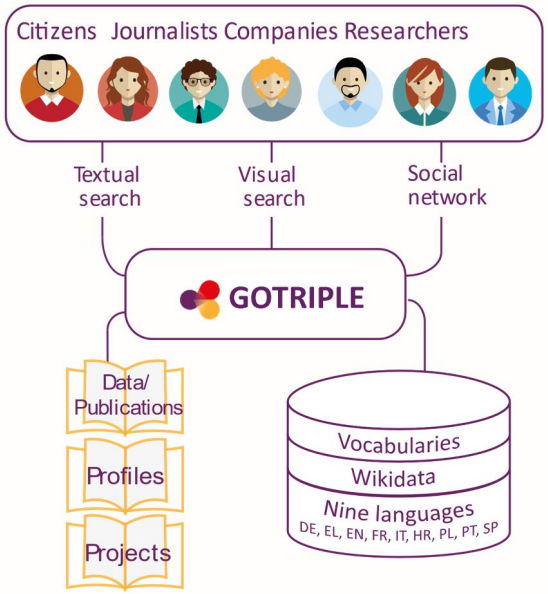
Triple

Transforming Research through Innovative Practices for Linked interdisciplinary Exploration

The SSH discovery platform GOTRIPLE: A future EOSC service

PRESENTERS
Francesca Di Donato (CNR)
Francesca.didonato@ilc.cnr.it
Stefanie Pohle (MWS)
pohle@maxweberstiftung.de

Max Weber Stiftung
Foundation German Humanities Institutes Abroad



WHY TRIPLE PROJECT?

- Strong fragmentation of SSH research
- SSH open scholarly resources (data, publications, other researchers' profiles and projects) currently scattered across local repositories
- Low use and reuse of SSH research
- Interdisciplinary collaboration possibilities are missed
- Societal impact is limited

www.gotriple.eu

TRIPLE will be a dedicated service of the OPERAS RI

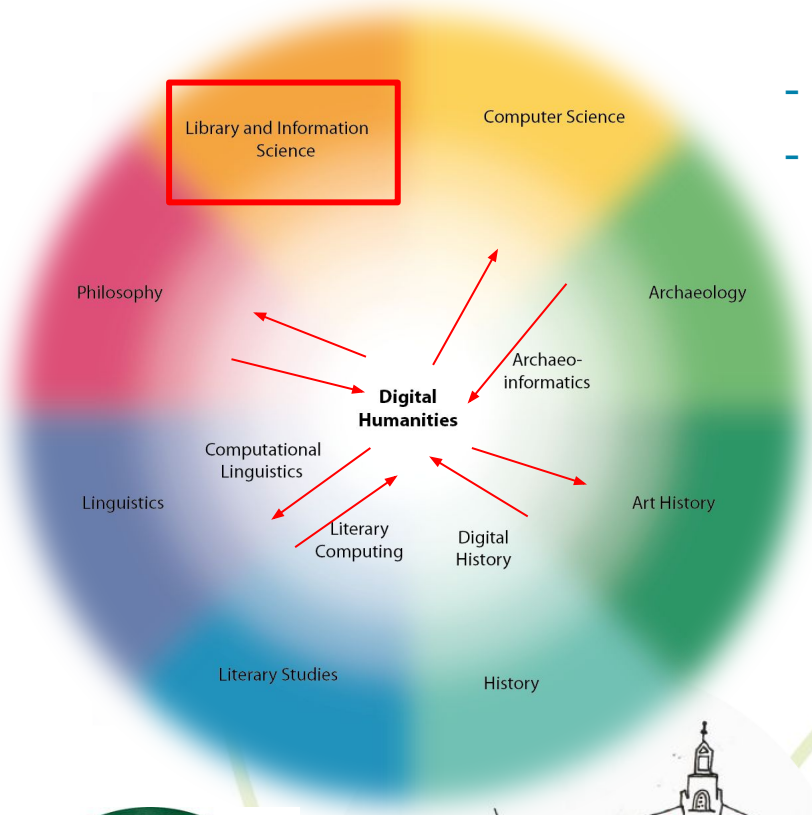
OPERAS
open access in the european research area through scholarly communication

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 863420.

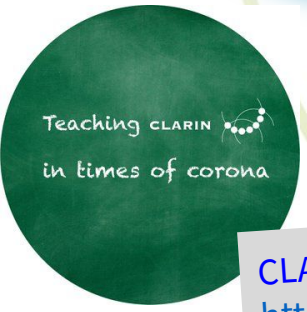
Construyendo las infraestructuras

- Las HD
- Infraestructuras y situación actual
- La importancia de las y los usuarios
- Servicios

Humanidades Digitales e infraestructuras



- Disciplinas: **Colaboración**
- Estudio de métodos/preguntas de investigación
  - **Método**: Cómo influyen las herramientas/recursos cuando se cuantifican algunos conceptos literarios (*Distant Reading*)
  - **Adaptar**: recursos de otros dominios
  - Modelado de datos, **metadatos**, bases de datos… (cómo se diseña el corpus literario y cuál es la pregunta de investigación)



CLARIN en el curriculum universitario:
https://labur.eus/scuzN

- Colaboración e interdisciplinaridad
  - Humanistas digitales
  - Críticos literarios
  - Lingüistas
  - Informáticos

Infraestructuras y red estratégica

- **INTELE**: red estratégica para la participación oficial en infraestructuras europeas CLARIN y DARIAH
  - **Impulsar** la investigación en humanidades y ciencias sociales
  - **Impulsar** los proyectos y programas **internacionales**
- **CLARIN: C**ommon **LA**nguage **R**esources and Technology **IN**frastructure
  - ESFRI ERIC (2012) y ESFRI Landmark (2016)
- **DARIAH: D**igital **R**esearch **I**nfrastructure for the **A**rts and **H**umanities
  - ESFRI ERIC (2014) y ESFRI Landmark (2016)

ESFRI European Strategy Forum on Research Infrastructures

ESFRI VISION

Equipping Europe with infrastructures for ground-breaking research

Research Infrastructures are strongly rooted in the regions and critically influence regional development

Horizon Europe will provide an opportunity to maximise the impact of Europe's Research Infrastructures.

The Mission initiative will promote the integration of the Research Infrastructure ecosystem whereby different RIs will cluster for a specific mission and develop joint services targeting complex research questions

https://www.esfri.eu/esfri-white-paper

INTELE: Red estratégica para la promoción de las **in**fraestructuras de **te**cnologías del **le**nguaje en eHumanidades y ciencias sociales

(1) **Impulsar actividades** de promoción de las infraestructuras CLARIN y DARIAH

(2) **Conectar grupos investigadores** que tengan interés para participar en dichas infraestructuras europeas

(3) Elaborar un **catálogo de herramientas y casos de uso** para castellano y lenguas cooficiales (euskera, catalán, gallego)

(4) Elaborar un informe para la **reevaluación positiva** de dichas infraestructuras

German Rigau
UPV/EHU

Núria Bel
UPF

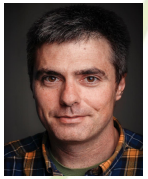
Dolores Romero
UCM

Manuel Gonzalez
ILG

M. Angel García
UJAEN

Borja Navarro
U. Alicante

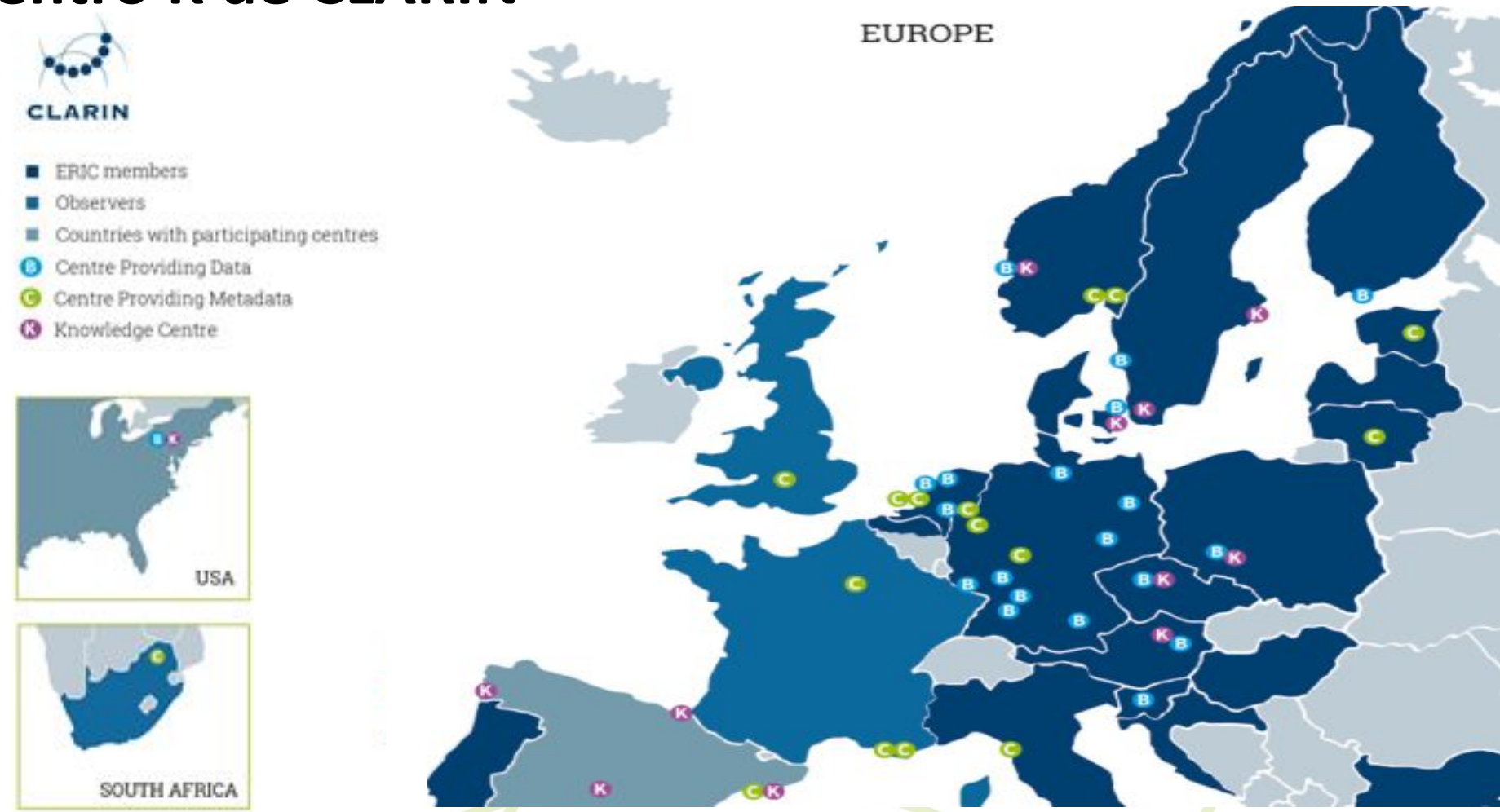
Salvador Ros
UNED



Mikel Iruskieta
UPV/EHU

Ainara Estarrona
UPV/EHU

Centro K de CLARIN



Núria Bel
UPF

Mikel Iruskieta
UPV/EHU

Salvador Ros
UNED

Xavier Gómez Guinovart UVIGO

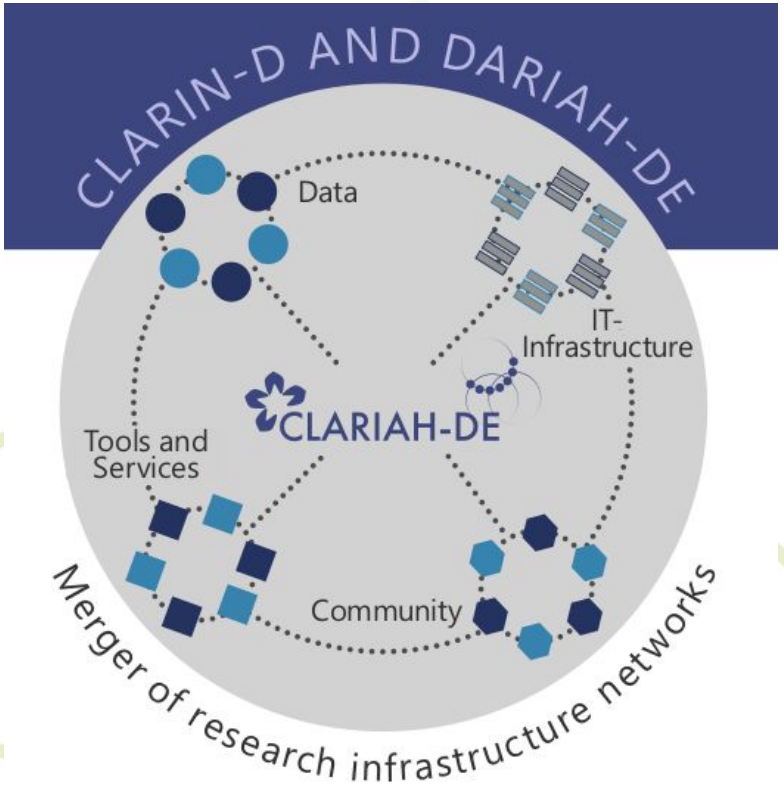
CLARIN CENTRE K

¿Qué es una infraestructura?

ESFRI European Strategy Forum on Research Infrastructures

THE NETWORK IN EUROPE

The CLARIN, DARIAH and CLARIAH research infrastructures are active on a national and the European level. CLARIN and DARIAH both have the status of a European Research Infrastructure Consortium (ERIC).



The content-related and technological foundations created by CLARIN-D and DARIAH-DE will be aligned, integrated, further developed and jointly maintained in CLARIAH-DE.

¿Por qué una infraestructura?

Investigar más rápido y con mayor calidad
- Tener más tiempo para investigar
  - - tiempo programando
  - - tiempo creando recursos
  - + impacto social
  - + reutilización…

TL y situación de las lenguas

- 5 clases de situaciones en **2012**
  - Situación excelente:
  - Situación adecuada: inglés
  - Situación media: alemán **castellano**, francés holandés …
  - Situación en desarrollo: **euskara**, **gallego**, **catalán**, esloveno, …
  - Situación pobre: irlandés (gaélico), islandés, rumano, …

- Proyecto Europeo EUROPEAN LANGUAGE EQUALITY (ELE) 2020
  - Desarrollar una agenda estratégica de investigación e innovación, y una hoja de ruta para lograr la igualdad total de las lenguas europeas en el ámbito digital para 2030.

Colaboración en la infraestructura: CLARIAH

National Roadmap for Large-Scale Research Infrastructure 1 General information

GENeric functionality, LINguistics, Socio-Economic History and Media Studies, CONtent of texts: history, literary

Fuente: https://www.clariah.nl/over/bestanden/downloads/send/10-folders/166-clariah-plus

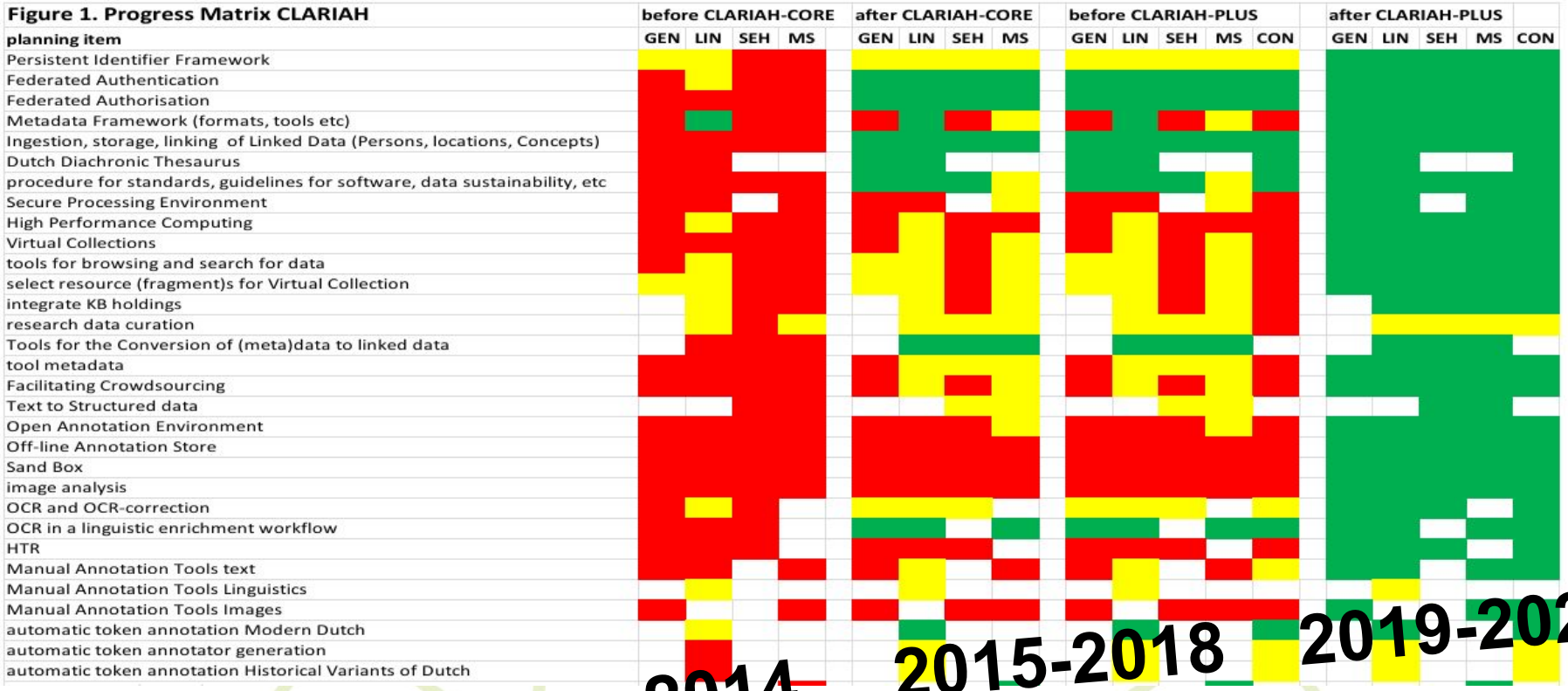


Figure 1. Progress Matrix CLARIAH	before CLARIAH-CORE				after CLARIAH-CORE				before CLARIAH-PLUS					after CLARIAH-PLUS				
planning item	GEN	LIN	SEH	MS	GEN	LIN	SEH	MS	GEN	LIN	SEH	MS	CON	GEN	LIN	SEH	MS	CON
Persistent Identifier Framework																		
Federated Authentication																		
Federated Authorisation																		
Metadata Framework (formats, tools etc)																		
Ingestion, storage, linking of Linked Data (Persons, locations, Concepts)																		
Dutch Diachronic Thesaurus																		
procedure for standards, guidelines for software, data sustainability, etc																		
Secure Processing Environment																		
High Performance Computing																		
Virtual Collections																		
tools for browsing and search for data																		
select resource (fragment)s for Virtual Collection																		
integrate KB holdings																		
research data curation																		
Tools for the Conversion of (meta)data to linked data																		
tool metadata																		
Facilitating Crowdsourcing																		
Text to Structured data																		
Open Annotation Environment																		
Off-line Annotation Store																		
Sand Box																		
image analysis																		
OCR and OCR-correction																		
OCR in a linguistic enrichment workflow																		
HTR																		
Manual Annotation Tools text																		
Manual Annotation Tools Linguistics																		
Manual Annotation Tools Images																		
automatic token annotation Modern Dutch																		
automatic token annotator generation																		
automatic token annotation Historical Variants of Dutch																		

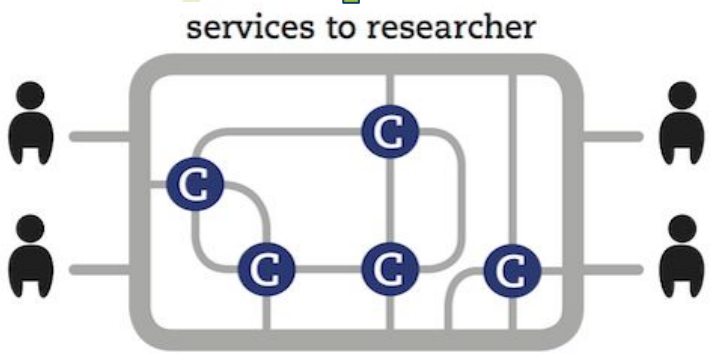
2014 2015-2018 2019-2023

CLARIN

Humanidades y Ciencias Sociales

- Desde una URL
  - **Los datos pueden estar en sitios diferentes**
- Datos sobre la lengua
  - Texto y vídeo
- Herramientas avanzadas e interoperables
  - buscar, analizar, combinar y crear...

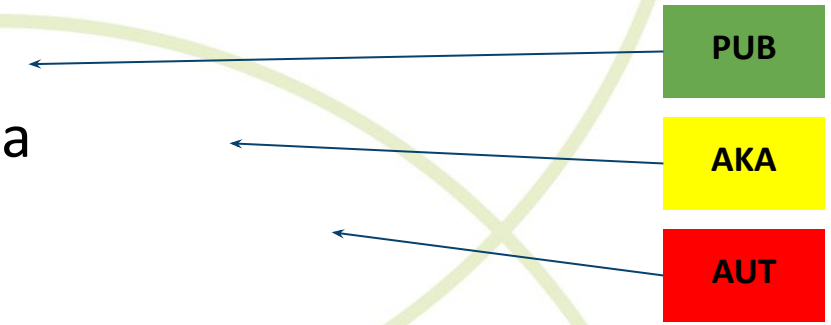
- **Facilitar el uso** de las TL



La infraestructura más que un proyecto

- Acceso confederado a todos los recursos y datos desde una única web
- **Estándares**
- **Protocolos** comunes
- Ayuda para el cambio de paradigma
- Diseño de los recursos **estratégicos**

- Corpora: abiertos y **púb**licos
- Abiertos solo para la **aca**demia
- Únicamente para **aut**orizados



CLARIN: experiencias de usuarios

https://zenodo.org/record/4288980#.X9YDS7N7mDI

Servicio ad hoc del centro CLARIN K

Publicado en el Tour de CLARIN

- Tesis en *Basque Center on Cognition, Brain and Language* (BCBL)
- Tema: “My PhD work focuses on the amount of exposure to each language within bilingual contexts, and how it shapes language acquisition at a cognitive and neural level”
  - **Herramientas**:
  - ANALHITZA
  - https://switchboard.clarin.eu

https://www.clarin.eu/blog/tour-de-clarin-interview-jose-perez-navarro

Language Resource Switchboard

Para encontrar la herramienta adecuada para tu tipo de datos lingüísticos

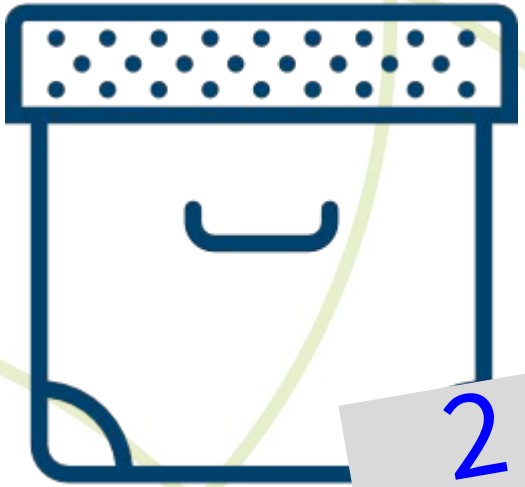
- https://switchboard.clarin.eu/

1

Depositing services

Para depositar y mantener corpora y recursos:

- www.clarin.eu/content/depositing-services



2

Language resources

Corpus y metadatos: en grandes cantidades y para búsquedas rápidas:

- vlo.clarin.eu/#tour
- contentsearch.clarin.eu
- https://labur.eus/gZ1ld

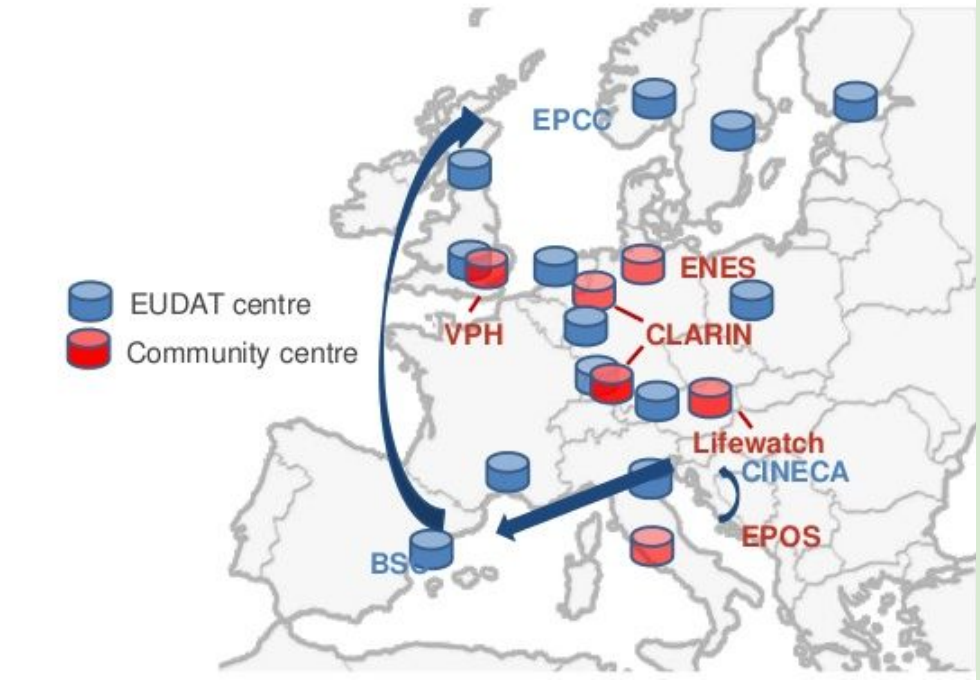
Virtual collections

Para crear corpus virtuales y poder mencionarlos (replicabilidad):

- https://www.clarin.eu/content/virtual-collections

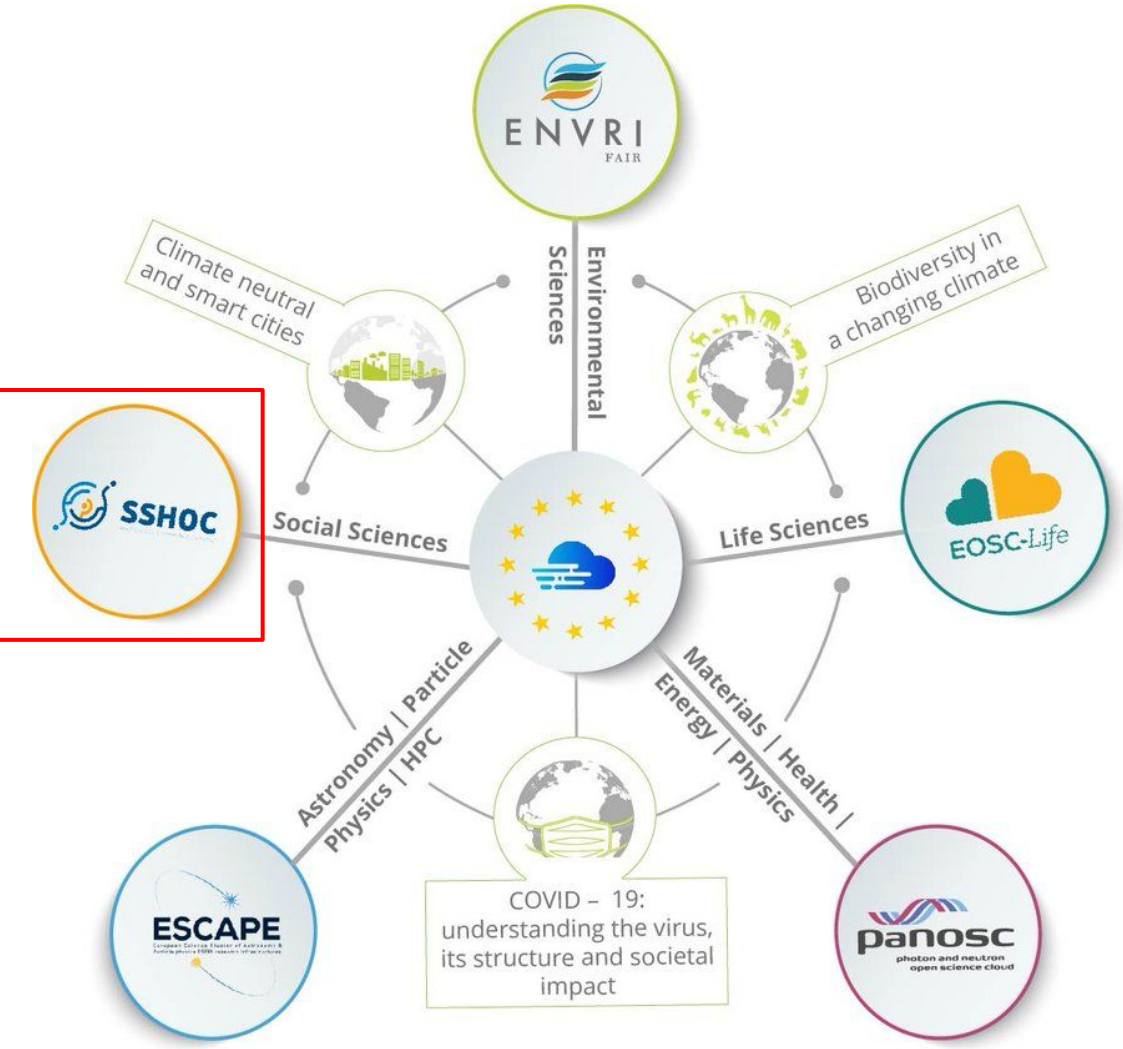
Ecosistema de investigación europeo

- EOSC
- SSHOC
- Interoperabilidad desde fuera



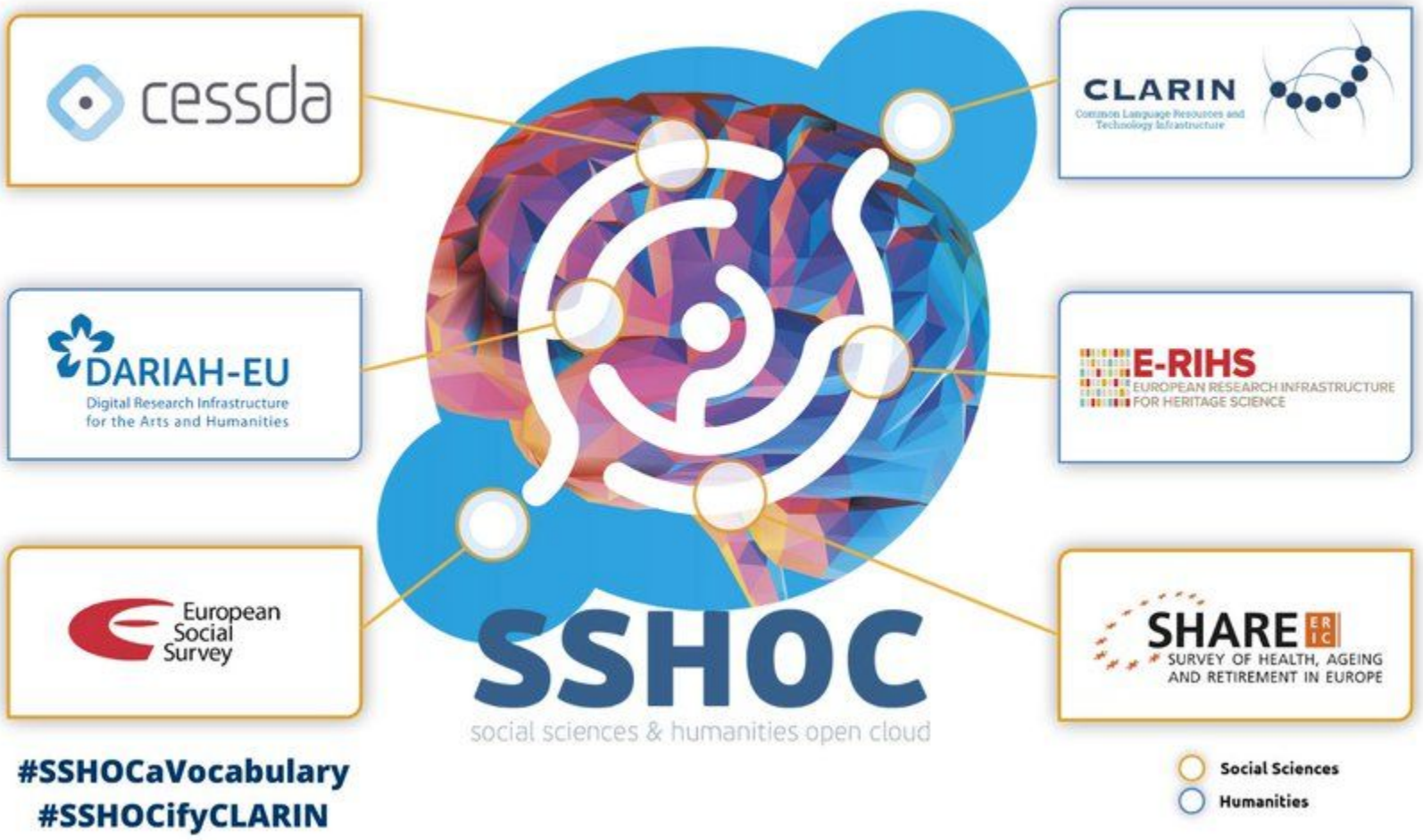
https://www.slideshare.net/EUDAT/b2-safe-how-to-replicate-your-data

Mapa de las infraestructuras EOSC



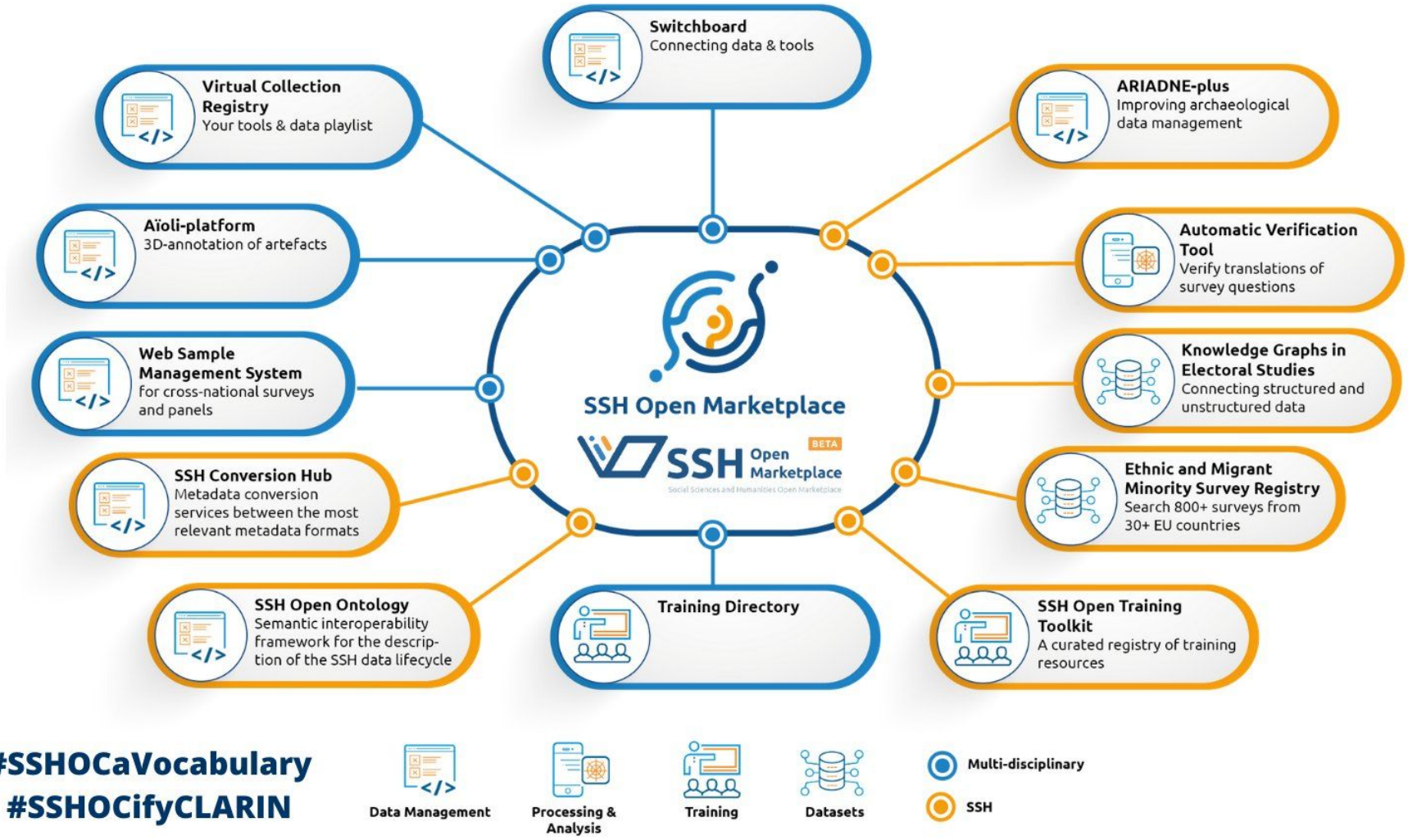
#SSHOC es una de las acciones de INFRAEOSC 04-2018, que consolida y conecta las e-infraestructuras europeas en ***European Open Science Cloud***.

SSHOC: Conexión de las e-infraestructuras europeas en HD y CCSS



SSH Open Marketplace y CLARIN

@SSHOpenCloud Objetivo: crear un escenario sostenible y perdurable para compartir y optimizar los datos y servicios en CCSS



Posibilidades del usuario

- Descubrir datos
- Repositorios de datos
- Archivos avanzados de datos
- Archivos del grupo de investigación o infraestructura
- Archivos personales
- Análisis de datos y procesamiento del lenguaje

Relación usuario/comunidad con infraestructura

- Se piden servicios y casos de uso a la comunidad
- Se evalúa y se ajusta la tecnología
- Se ofrecen los servicios a la comunidad

Creando mi corpus virtual de Educación Infantil en CLARIN Virtual collections

Virtual Collection Registry | Browse | Create | My Collections | Help | mikel.iruskieta_ehu.eus@clarin.eu | CLARIN

Euskarazko haurren corpusak

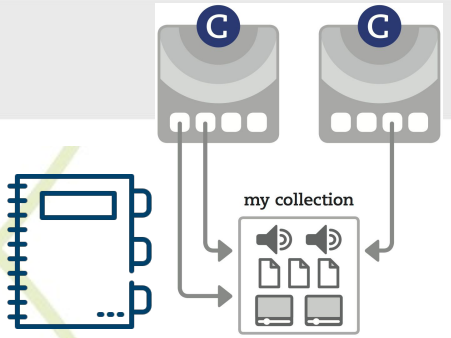
General

Name: Euskarazko haurren corpusak

Resources

Reference	Actions
Frogs French Iduguine Corpus	•••
Basque SotoValle Corpus - 040505	•••
Basque Luque Corpus - 33cas3	•••
Haur Hezkuntzako ipuin-bilduma HDL 11304/f27f5e92-af01-4a37-a6d9-82cf14afa160	

State	Type	Created
private	extensional	2021-06-29



Obtener y transformar: casos de uso digitalización: herramientas y servicios en CLARIN

¿Qué hacer con textos digitalizados?

- Fondos digitales de bibliotecas
- Tu libro a estudiar digitalizado
- Centro de competencia IMPACT
- Transkribus

IMPACT CLARIN K-centre y BNE

impact
digitisation.eu
centre of competence

Vida de Lazarillo de Tormes

IMPACT DATASET BROWSER

This resource is property of BIBLIOTECA NACIONAL DE ESPAÑA BNE

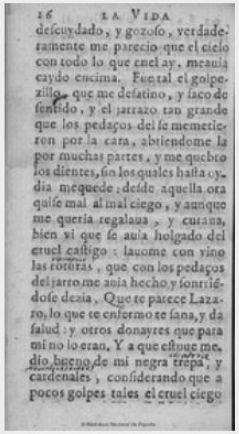
and distributed by the Impact Centre of Competence

impact
digitisation.eu
centre of competence

ABOUT ▾ TOOLS & RESOURCES ▾ KNOWLEDGE ▾ COMMUNITY ▾ CKC HELPDESK

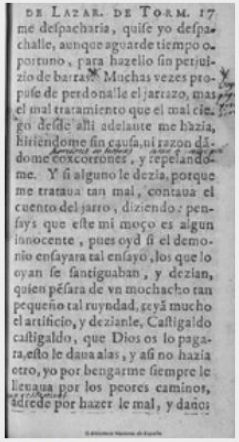
440435

TIFF XML



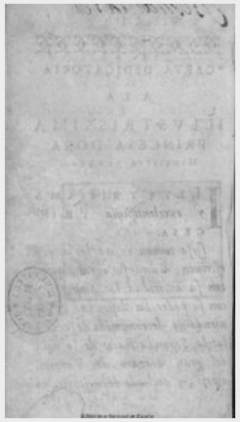
440436

TIFF XML



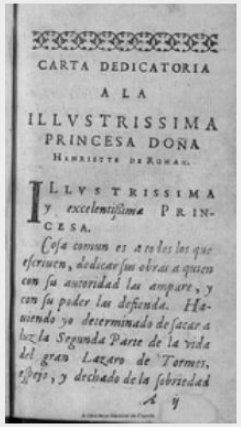
440437

TIFF XML



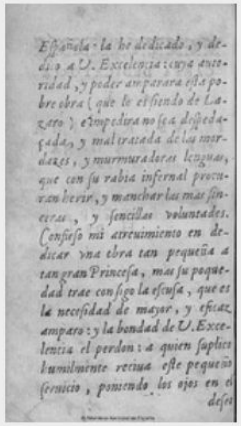
440438

TIFF XML



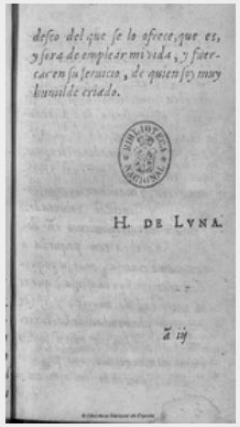
440439

TIFF XML



440440

TIFF XML

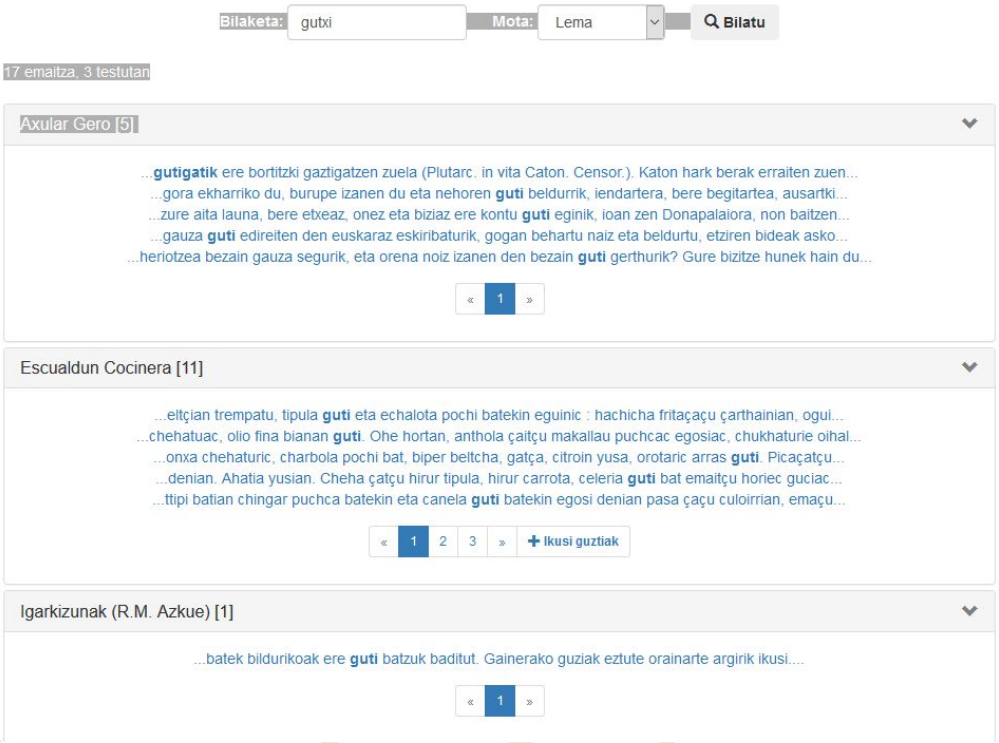


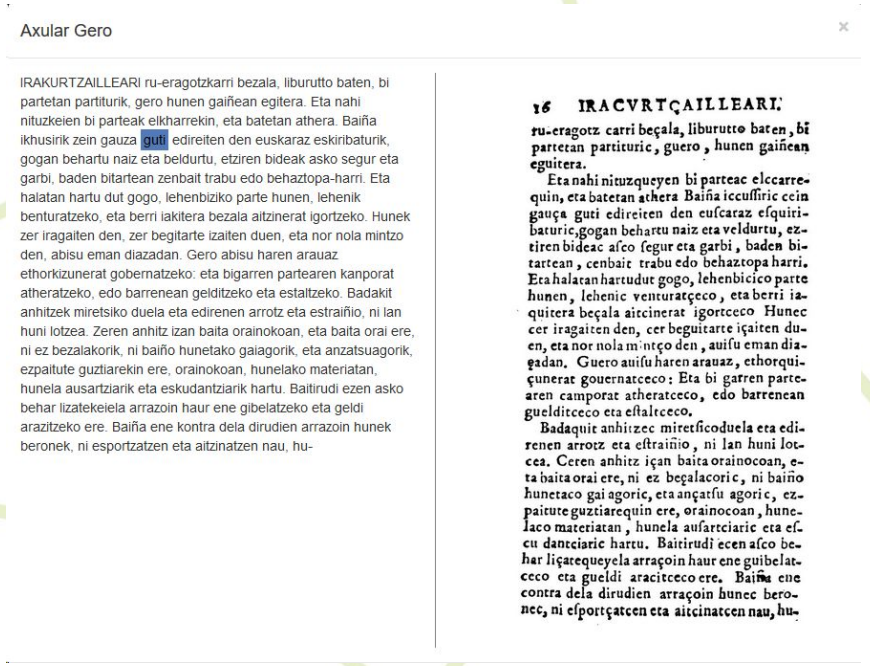
440441

TIFF XML

Digitalizar y normalizar euskera antiguo

Sintaktikoki etiketatutako euskarazko corpus historikoa





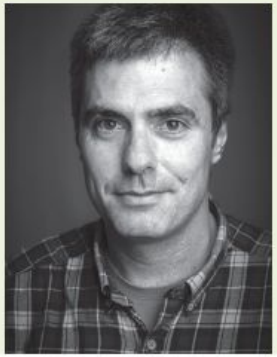
http://ixa2.si.ehu.eus/armiarma-bim2021-2/

http://ixa2.si.ehu.es/testuhistorikoak/

Uso del centro CLARIN-K IMPACT-CKC

IMPACT-CKC K-CENTRE

Interview | **Mikel Iruskieta**



Mikel Iruskieta is a computational linguist who is part of the Ixa Research Group and the Didactics of Language and Literature Department at the University of the Basque country. He has collaborated with the CLARIN IMPACT-CKC Knowledge Centre, which helped him and his colleagues digitize Basque texts.

Could you briefly describe your academic and research background?

<

My current research focuses on the didactics and analysis of Basque, mostly regarding discourse parsing and evaluation of discourse structure. For the last five years, I have mainly worked on adapting language technologies for teaching and learning purposes. With that goal, I have created and now co-lead a postgraduate programme in Basque (University Specialist in ICT and Digital Competences in Education, Continuing Education and Language Teaching), a research group working in Digital Humanities and Education. Our aim is to build a research community that will conduct research and teach in Basque by adopting a critical approach and using language technologies in a pedagogical context. In this postgraduate programme, my colleagues and I are developing a new framework of the socio-tech pedagogy for Basque that will cover the following topics:

- The Basics of Technology and Pedagogy;
- Formal Education and Technology;
- Continuing Education and Technology;
- Language Teaching and Technology Development;
- Society and Education, Opportunities and Risk of Technology;
- E-learning: Approaches and resources; and
- Digital Research: Methods and resources.

>

Does the fact that Basque is a language isolate have any bearing on the development of language tools tailored to it?

<

The history and current situation of the Basque language are both complex and interesting. Basque has a relatively small community of speakers (751,700 active and 1,185,500 passive speakers) which lives in contact with three powerful language communities, namely Spanish and French (as official languages in the Basque Country) and English (as a foreign language). It is also not supported enough by official language policies. As a result, Basque is still considered an under-resourced language. In this context, the work of the Ixa Group for NLP is highly valuable. They have developed basic resources for Basque (as well as for other languages) which are used by the research community, for example IXApipes (a modular set of NLP tools which provide easy access to NLP technology for several languages that can be used or exploit its modularity to pick and change different components) and ANALHITZA (a web service to analyse Basque, Spanish and English texts without needing any technical experience). Many more basic and advanced tools and resources for Basque can be found on the website of the HiTZ: Basque Center for Language Technology.

>

How did you get involved with the IMPACT K-Centre and how did they help you with your research?

<

I learned about the IMPACT K-Centre when they joined CLARIN. Because I was working on several different digitization projects for Basque and for Spanish, I immediately got in touch with them and asked for their help. Isabel Martínez Sempere, the manager of IMPACT, helped me solve a digitization issue that I encountered when I was analysing the most frequently occurring words in *Pulgarcito*, which is a Cuban children's magazine written in Spanish from 1919 to 1920. This magazine consists of very diverse materials, such as drawings and handwritten texts, which are

http://clarin-es.org/tour-de-clarin-vol-iii/

Digitize in Basque

Bazen bitan amarekin bizi zen neskatila bat, mendi, aintzira eta ibai emaritsuez inguratutako herrixka batean. Leku eder eta lasaia zen; bizitzak gorabehera eta kezka handirik gabe egiten zuen aurrera. Etxetik hurbil, zuhaitz hostotsuez eta kolore askotako basaloreez betetako baso handi bat zegoen. Zalantzarik gabe, huraxe zen txango bat egiteko zuen lekurik gogokoena. Halaxe egiten zuen amonak josi zion txango gorri batekin. Ez zuen sekula kentzen, bainu bat hartzeko eta lo egiteko soilik... eta amak behin eta berriz esaten zion ezin zuela hura jantzita zuela oheratu.

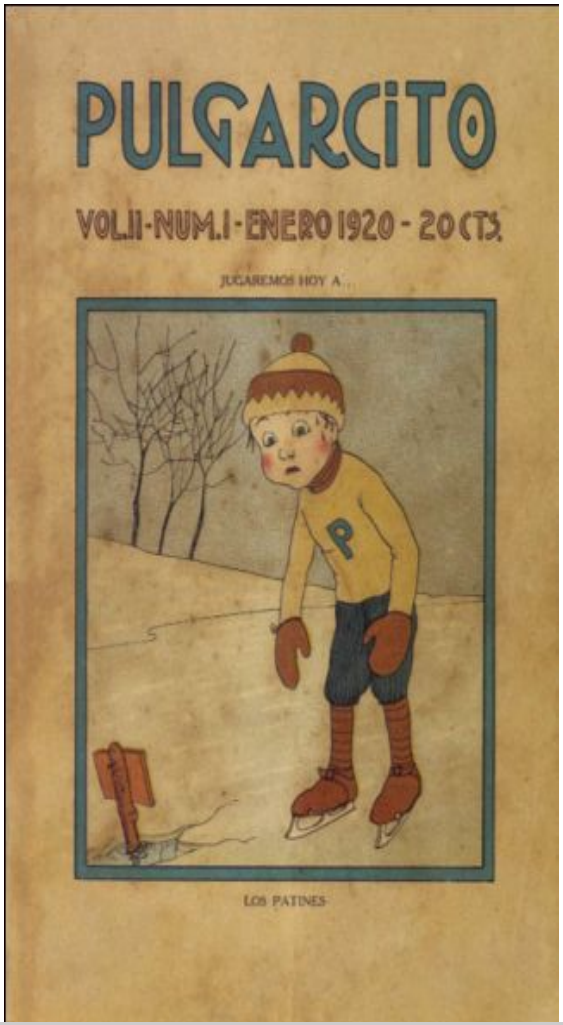
^f^azen bitan amarekin bizi zen neskatila bat, aintzira eta ibai emaritsuez inguratutako herrixka Leku eder eta lasaia zen; bizitzak gorabehera eta kezka handirikgabe egiten zuen aurrera. Etxetik hurbil, zuhaitz hostotsuez eta kolore askotako basaloreez betetako baso handi bat zegoen. Zalantzarik gabe, huraxe zen txango bat egiteko zuen lekurik gogokoena. Halaxe egiten zuen amonak josi zion txango gorri batekin. Ez zuen sekula kentzen, bainu bat hartzeko eta lo egiteko soilik... eta amak behin eta berriz esaten zion ezin zuela hura jantzita zuela oheratu.

CUATRO TUERCAS [EDT]

Ed. CUATRO TUERCAS:
http://www.cuatrotuercas.com/

impact
digitisation.eu
centre of competence

IMPACT CLARIN K-centre:
https://www.digitisation.eu/



Libro en castellano

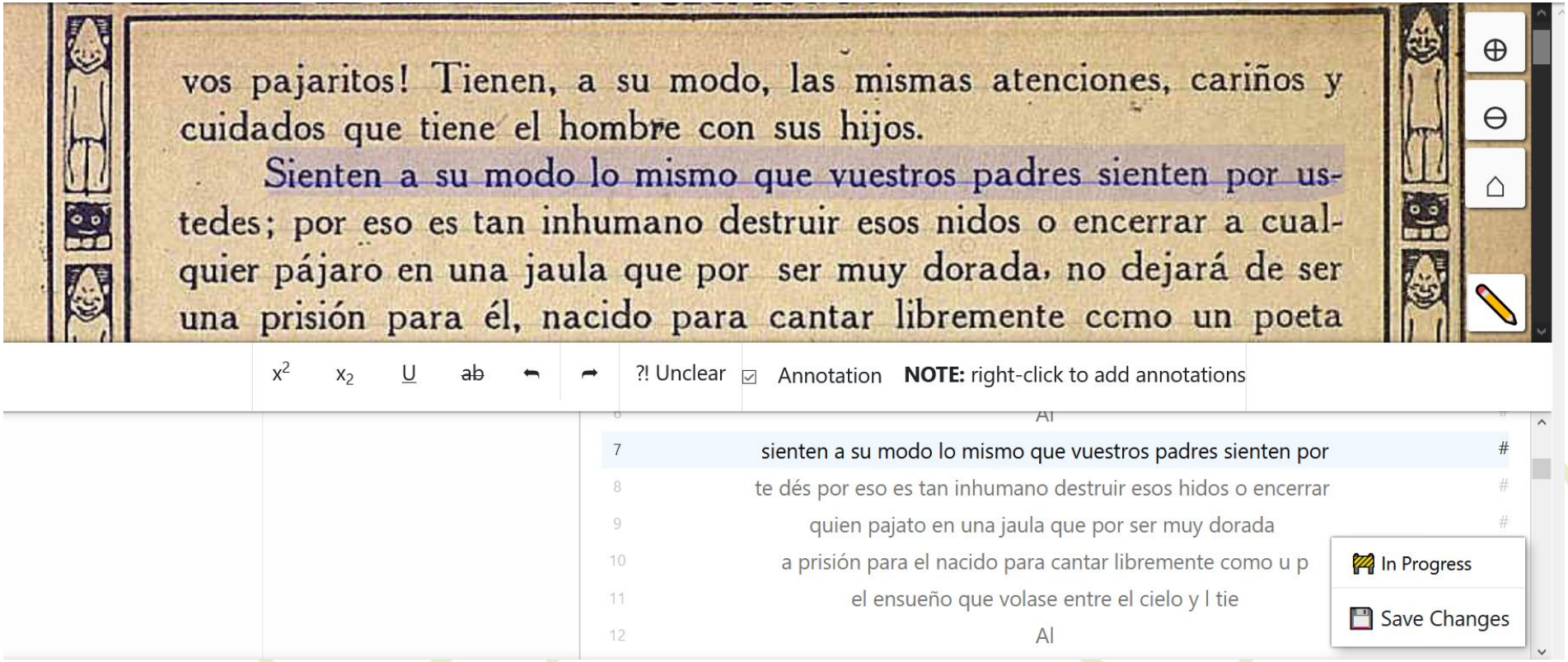
Objetivo: analizar un libro digitalizado con texto escrito a mano y a máquina

- Publicado en CLARIN
- Descarga: wget imagenes.sld.cu/download/pulgarcito/volumen-2.pdf

Interview at CLARIN: https://www.clarin.eu/blog/what-impact-k-centre-can-do-you

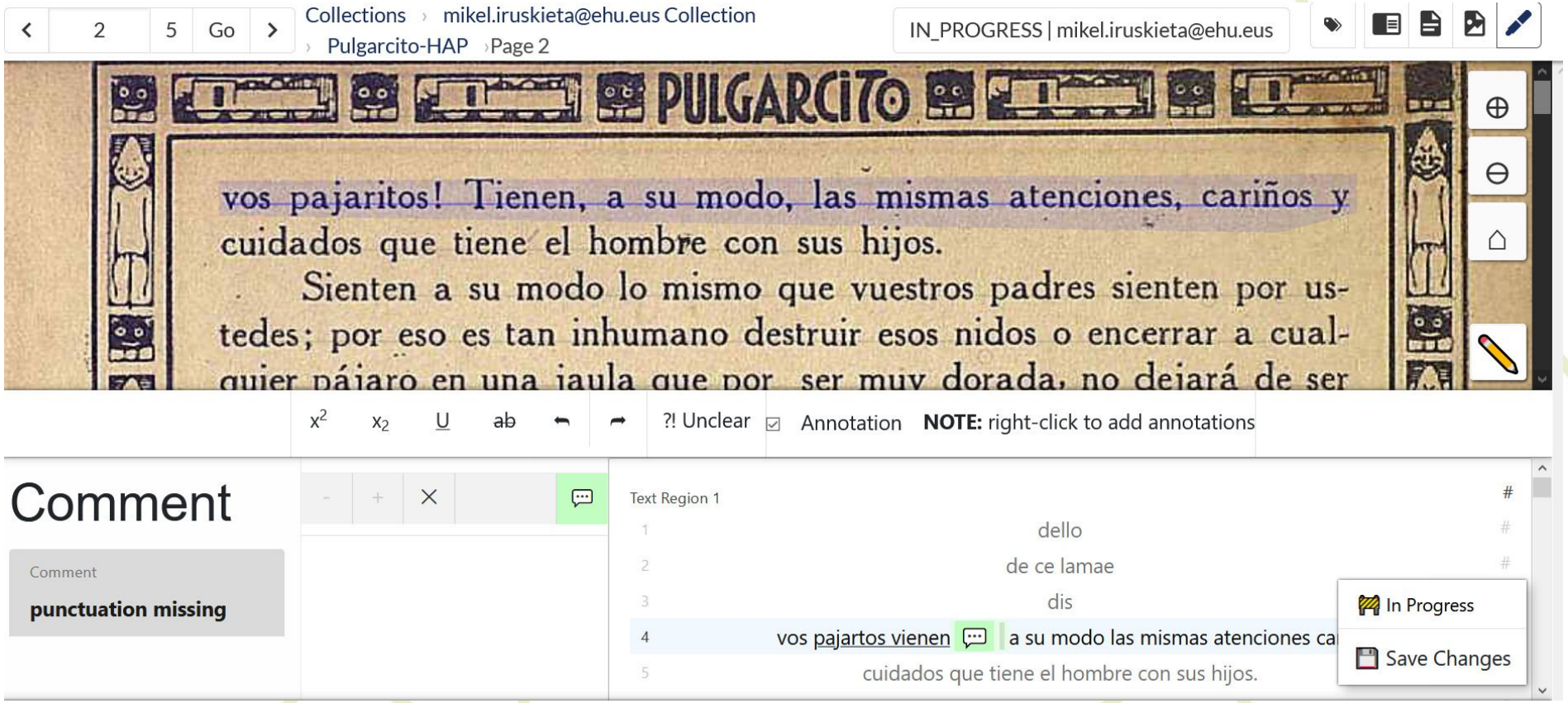
Más en [PH] Automated Downloading with Wget

Otra opción: Transcribir texto con Transcribus



https://transkribus.eu/lite/collection/89934/doc/617293/detail/2

Editar el texto transcrito

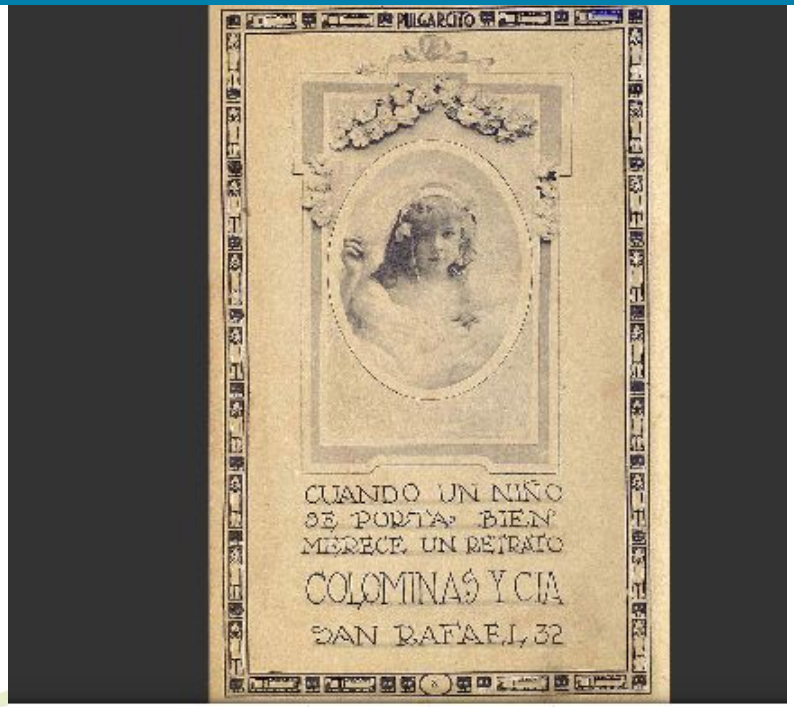


Ejemplo de texto manuscrito PDF2TXT (IMPACT)

CUANDO UN N1NO <<SI POEAA? M$&ECE, UMBETRAfO conminas y ca	

Comparación entre IMPACT vs Transcribus

CUANDO UN N1NO <<SI POEAA? M$&ECE, UMBETRAfO conminas y ca	c d Di d d ded le des
IMPACT	**Transcribus**

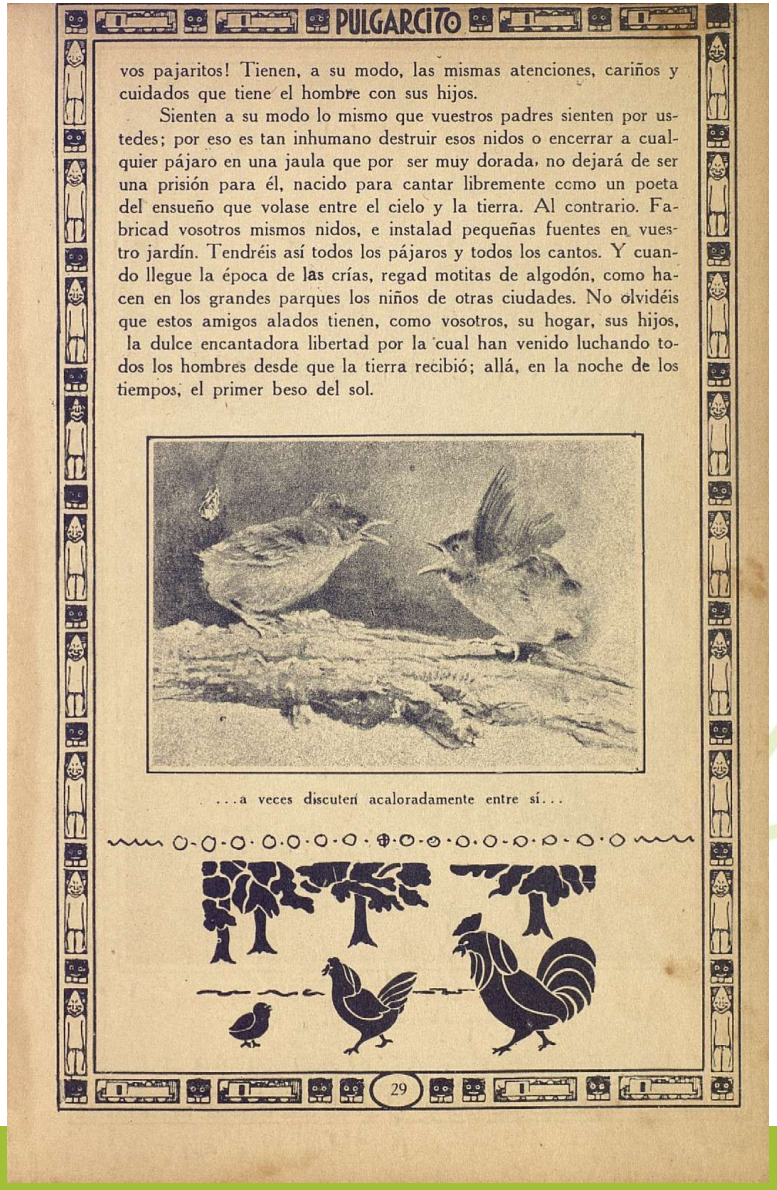


U ab ⮌ ⮎ ?! Unclear

c
d
Di
d d ded le
des

https://transkribus.eu/lite/collection/89934/doc/617293/detail/5

PDF2TXT (IMPACT)



PULGARCITO

vos pajaritos! Tienen, a su modo, las mismas atenciones, cariños y cuidados que tiene el hombre con sus hijos.

Sienten a su modo lo mismo que vuestros padres sienten por ustedes; por eso es tan inhumano destruir esos nidos o encerrar a cualquier pájaro en una jaula que por ser muy dorada, no dejará de ser una prisión para él, nacido para cantar libremente como un poeta del ensueño que volase entre el cielo y la tierra. Al contrario. Fabricad vosotros mismos nidos, e instalad pequeñas fuentes en vuestro jardín. Tendréis así todos los pájaros y todos los cantos. Y cuando llegue la época de las crías, regad motitas de algodón, como hacen en los grandes parques los niños de otras ciudades. No olvidéis que estos amigos alados tienen, como vosotros, su hogar, sus hijos, la dulce encantadora libertad por la cual han venido luchando todos los hombres desde que la tierra recibió; allá, en la noche de los tiempos, el primer beso del sol.

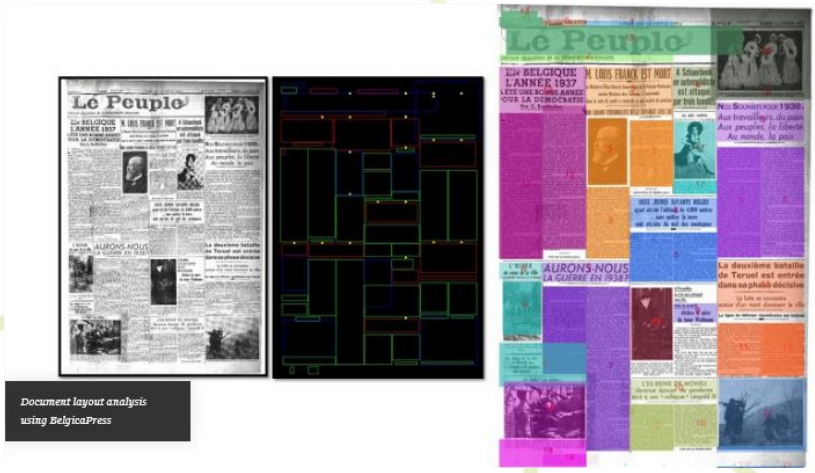
...a veces discuten acaloradamente entre sí...

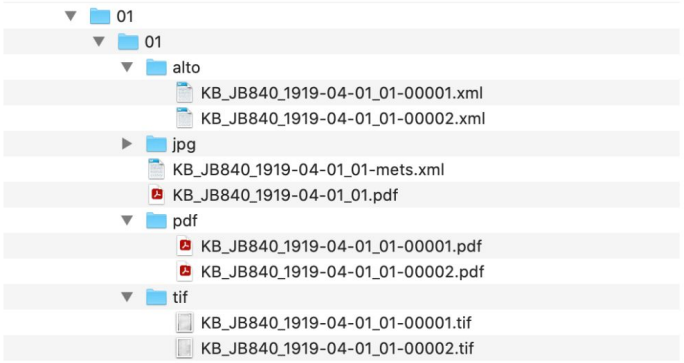
29

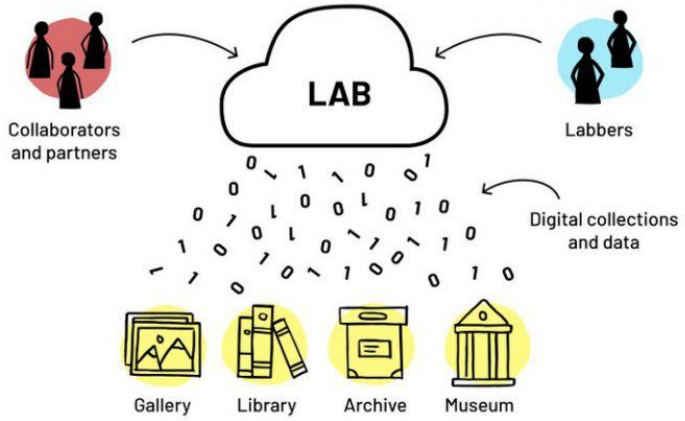
vos pajaritos! Tienen, a su modo, las mismas atenciones, cariños y cuidados que tiene el hombre con sus hijos. Sienten a su modo lo mismo que vuestros padres sienten por us- tedes; por eso es tan inhumano destruir esos nidos o encerrar a cual- quier pajaro en una jaula que por ser muy dorada, no dejara de ser una prision para el, nacido para cantar libremente ccmo un poeta del ensueño que volase entre el cielo y la tierra. Al contrario. Fa- bricad vosotros mismos nidos, e instalad pequeñas fuentes en vues- tro jardn. Tendreis as todos los pajaros y todos los cantos. Y cuan- do llegue la epoca de las cras, regad motitas de algodon, como ha- cen en los grandes parques los niños de otras ciudades. No olvideis que estos amigos alados tienen, como vosotros, su hogar, sus hijos, la dulce encantadora libertad por la cual han venido luchando to- dos los hombres desde que la tierra recibio; alla, en la noche de los tiempos, el primer beso del sol.	vos pajaritos! Tienen, a su modo, las mismas atenciones, cariños y 4 cuidados que tiene el hombre con sus hijos. Sienten a su modo lo mismo que vuestros padres sienten por us- tedes; por eso es tan inhumano destruir esos nidos o encerrar a cual- quier pajaro en una jaula que por ser muy dorada, no dejara de ser una prision para el, nacido para cantar libremente ccmo un poeta del ensueño que volase entre el cielo y la tierra. Al contrario. Fa- bricad vosotros mismos nidos, e instalad pequeñas fuentes en. vues- tro jardín. Tendreis as todos los pajaros y todos los cantos. Y cuan- do llegue la epoca de las cras, regad motitas de algodon, como ha- cen en los grandes parques los niños de otras ciudades. No olvideis que estos amigos alados tienen, como vosotros, su hogar, sus hijos, la dulce encantadora libertad por la `cual han venido luchando to- dos los hombres desde que la tierra recibio; alla, en la noche de los tiempos, el primer beso del sol. ...a veces discuten acaloradamente entre s... O-O-O-O'O'O-O - $-0.0-0.0-0-0 -
NoBin pul-074.txt	Bin pul-074.txt

DATA-KBR-BE: data as collection. DARIAH

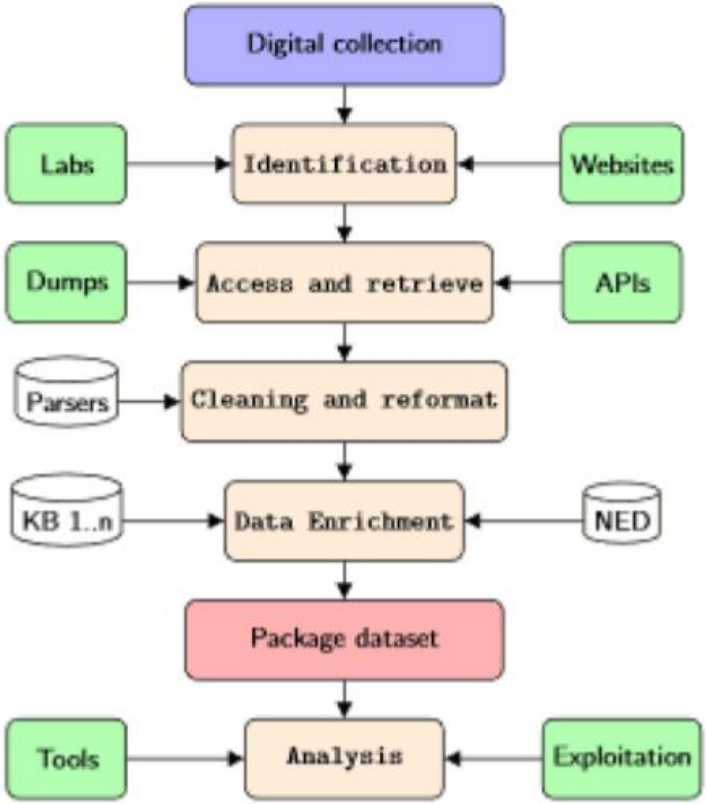
- **Facilitar datos y crear ediciones digitales para investigación de HD**
  - Diseñar el flujo de trabajo para la extracción de datos adecuados
  - Diseñar la plataforma Open Data
  - Inventario de colecciones digitales
  - Publicación de los datasets
  - Hackathon usando los datasets







Data as collection. BVMC



Workshop INTELE: "Facilitando el acceso computacional a colecciones digitales". (Biblioteca Virtual Miguel de Cervantes)

Sesión práctica: github.com/hibernator11/notebook-ph

SSHOC: Train–the-Trainer Bootcamp for Librarians: https://zenodo.org/record/3970799#.YMn2gqZ7mL1

Analizar y presentar: Casos de uso análisis textual en CLARIN

- Interoperabilidad y el análisis textual
- Interoperabilidad y la transcripción de videos
- Otros casos de uso:
  - Análisis textual y mapas
  - Historias de usuarios
  - Survival kit

Uso de las infraestructuras

Texto escrito:

1. Eudat: FAIR data across borders and disciplines
2. Interoperable con el Switchboard de CLARIN
3. Detección de lengua y propone herramientas/métodos
4. Analiza el texto y ofrece técnicas de visualización

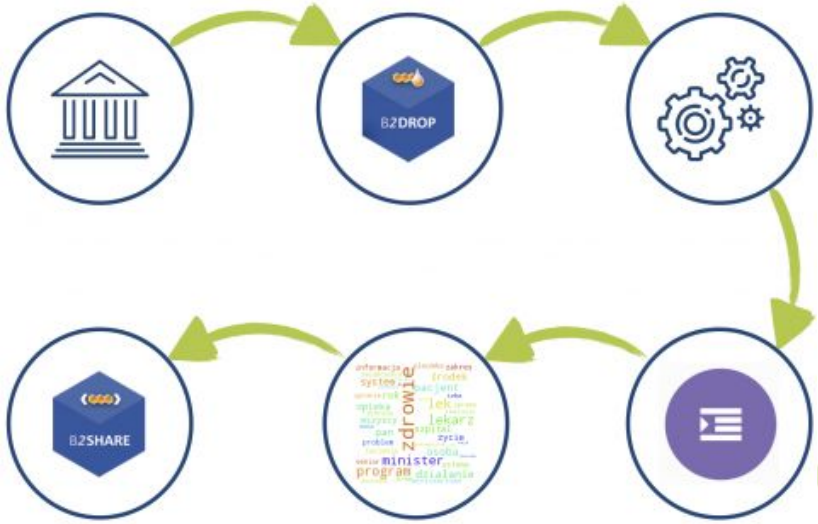
De voz a texto:

5. Servicio BAS de CLARIN
6. Decenas de lenguas y variaciones
7. Múltiples formatos de salida para seguir con la investigación: TXT, SCV, PRAAT, Video...

Investigar en la nube en CLARIN: principios FAIR

Ejemplo en euskera en la EOSC y EuDAT

- Datos:
  - Corpus de cuentos infantiles
- Análisis sintáctico
- Análisis en la nube
- Publicación persistente



Interoperable

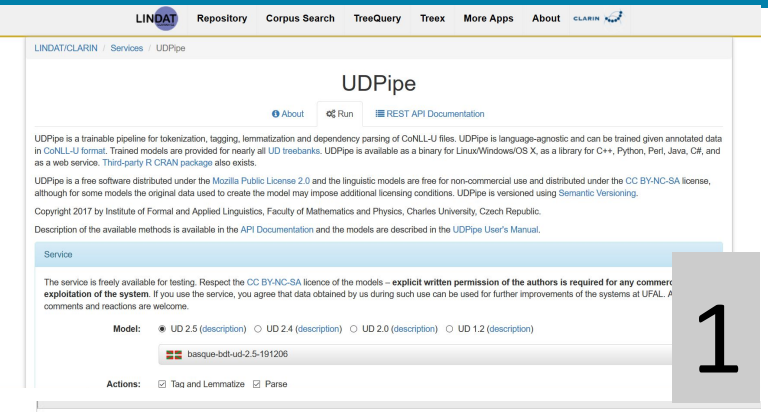
Original en inglés:
https://www.clarin.eu/showcase/eosc-portal-demonstration

Recursos en la nube para texto

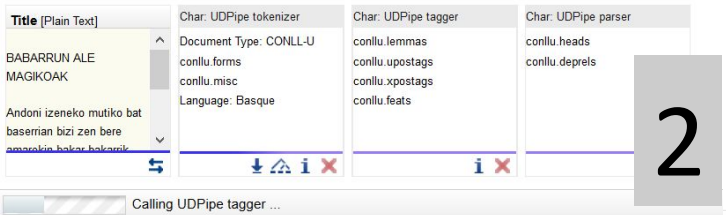
1. Corpus en Eudat
2. Analizar con un clic en Switchboard
3. Elegir un recurso para el análisis de texto

- Recursos
  - Para el castellano: 4
  - Para el inglés: 16
  - Para el aleman: 13
  - Para el polaco: 26

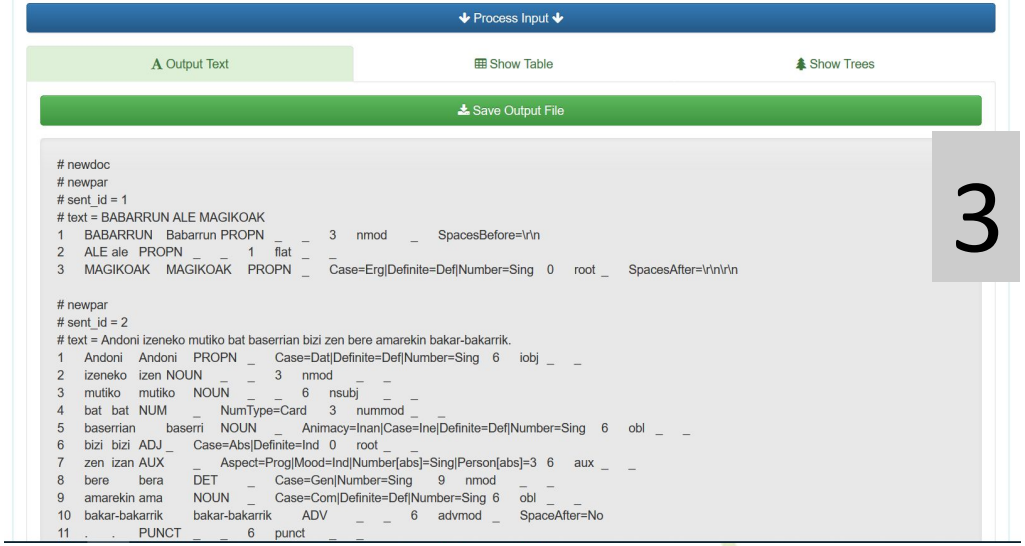
1. Constituency Parsing
2. Coreference Resolution
3. Dependency Parsing
4. Distant Reading
5. Extraction of Polish terminology
6. Inclusion detection
7. Keyword Extractor
8. Lemmatization
9. Machine Translation
10. Metadata Processing
11. Morpho-syntactic tagger
12. Morphological Analysis
13. Named Entity Recognition
14. Named Entity Relation Detection
15. Part-Of-Speech Tagging
16. Sentiment Analysis
17. Shallow Parsing
18. Spatial expression detection
19. Speech Recognition
20. Stylometry
21. TF, IDF, TF-IDF calculation
22. Text Analytics
23. Text Enhancement
24. Text Summarization
25. Tokenisation
26. Topic Modelling
27. Visualisation of Geographic Data
28. Word sense disambiguation


LINDAT
Repository
Corpus Search
TreeQuery
Treex
More Apps
About
LINDAT/CLARIN / Services / UDPipe
UDPipe
About
Run
REST API Documentation
UDPipe is a trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. UDPipe is language-agnostic and can be trained given annotated data in CoNLL-U format. Trained models are provided for nearly all UD treebanks. UDPipe is available as a binary for Linux/Windows/OS X, as a library for C++, Python, Perl, Java, C#, and as a web service. Third-party R CRAN package also exists.
UDPipe is a free software distributed under the Mozilla Public License 2.0 and the linguistic models are free for non-commercial use and distributed under the CC BY-NC-SA license, although for some models the original data used to create the model may impose additional licensing conditions. UDPipe is versioned using Semantic Versioning.
Copyright 2017 by Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Czech Republic.
Description of the available methods is available in the API Documentation and the models are described in the UDPipe User's Manual.
Service
Model:
UD 2.5 (description)
UD 2.4 (description)
UD 2.0 (description)
UD 1.2 (description)
basque-bdt-ud-2.5-191206
Actions:
Tag and Lemmatize
Parse

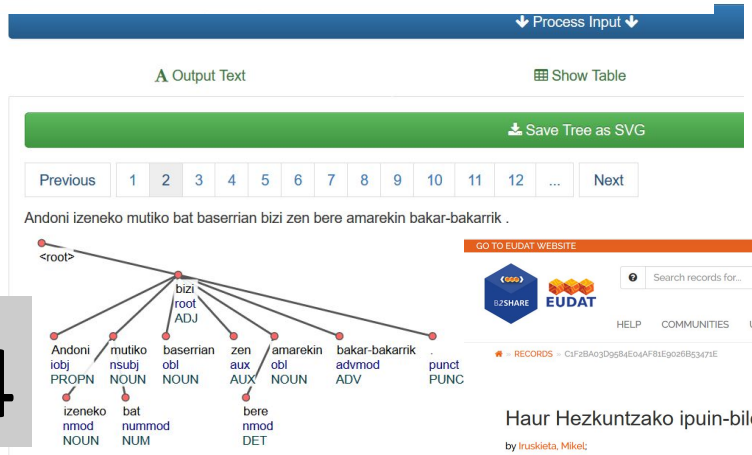
1


Title [Plain Text]
BABARRUN ALE MAGIKOAK
Andoni izeneko mutiko bat baserrian bizi zen bere
Char: UDPipe tokenizer
Document Type: CONLL-U
conllu.forms
conllu.misc
Language: Basque
Char: UDPipe tagger
conllu.lemmas
conllu.upostags
conllu.xpostags
conllu.feats
Char: UDPipe parser
conllu.heads
conllu.deprels
Calling UDPipe tagger ...

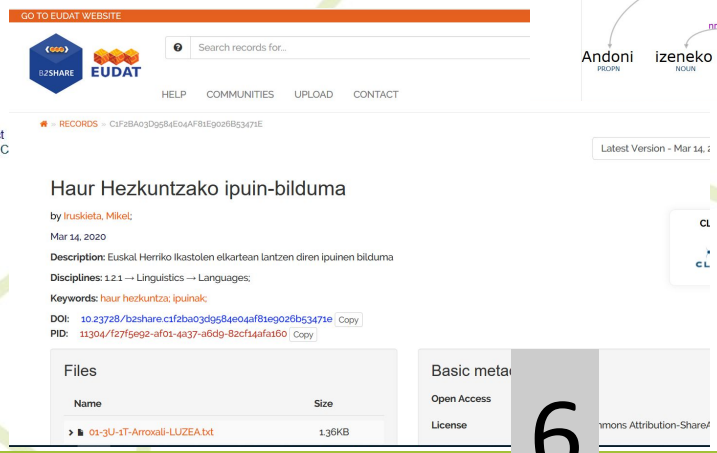
2


Process Input
Output Text
Show Table
Show Trees
Save Output File
newdoc
newpar
sent_id = 1
text = BABARRUN ALE MAGIKOAK
1 BABARRUN Babarrun PROPN _ _ 3 nmod _ SpacesBefore=\r\n
2 ALE ale PROPN _ _ 1 flat _ _
3 MAGIKOAK MAGIKOAK PROPN _ Case=Erg|Definite=Def|Number=Sing 0 root _ SpacesAfter=\r\n\r\n
newpar
sent_id = 2
text = Andoni izeneko mutiko bat baserrian bizi zen bere amarekin bakar-bakarrik.
1 Andoni Andoni PROPN _ Case=Dat|Definite=Def|Number=Sing 6 iobj _ _
2 izeneko izen NOUN _ _ 3 nmod _ _
3 mutiko mutiko NOUN _ _ 6 nsubj _ _
4 bat bat NUM _ NumType=Card 3 nummod _ _
5 baserrian baserri NOUN _ Animacy=Inan|Case=Ine|Definite=Def|Number=Sing 6 obl _ _
6 bizi bizi ADJ _ Case=Abs|Definite=Ind 0 root _ _
7 zen izan AUX _ Aspect=Prog|Mood=Ind|Number[abs]=Sing|Person[abs]=3 6 aux _ _
8 bere bera DET _ Case=Gen|Number=Sing 9 nmod _ _
9 amarekin ama NOUN _ Case=Com|Definite=Def|Number=Sing 6 obl _ _
10 bakar-bakarrik bakar-bakarrik ADV _ _ 6 advmod _ SpaceAfter=No
11 . . PUNCT _ _ 6 punct _ _

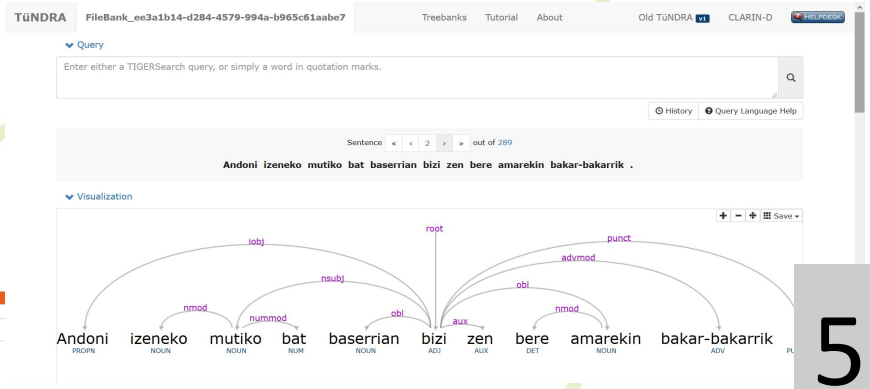
3


Process Input
Output Text
Show Table
Save Tree as SVG
Previous
1 2 3 4 5 6 7 8 9 10 11 12 ...
Next
Andoni izeneko mutiko bat baserrian bizi zen bere amarekin bakar-bakarrik .
<root>
bizi root ADJ
Andoni iobj PROPN
mutiko nsubj NOUN
baserrian obl NOUN
zen aux AUX
amarekin obl NOUN
bakar-bakarrik advmod ADV
. punct PUNC
izeneko nmod NOUN
bat nummod NUM
bere nmod DET

4


GO TO EUDAT WEBSITE
B2SHARE
EUDAT
Search records for..
HELP
COMMUNITIES
UPLOAD
CONTACT
RECORDS
Haur Hezkuntzako ipuin-bilduma
by Iruskieta, Mikel;
Mar 14, 2020
Description: Euskal Herriko Ikastolen elkartean lantzen diren ipuinen bilduma
Disciplines: 12.1 → Linguistics → Languages;
Keywords: haur hezkuntza; ipuinak;
DOI: 10.23728/b2share.c1f2ba03d9584e04af81e9026b53471e Copy
PID: 11304/f27f5e92-af01-4a37-a6d9-82cf14afa160 Copy
Files
Name
Size
01-3U-1T-Arroxali-LUZEA.txt
1.36KB
Basic meta
Open Access
License

6


TÜNDRA
FileBank_ee3a1b14-d284-4579-994a-b965c61aabe7
Treebanks
Tutorial
About
Old TüNDRA v1
CLARIN-D
HELPDESK
Query
Enter either a TIGERSearch query, or simply a word in quotation marks.
History
Query Language Help
Sentence « ‹ 2 › » out of 289
Andoni izeneko mutiko bat baserrian bizi zen bere amarekin bakar-bakarrik .
Visualization
Save
iobj
root
punct
advmod
nsubj
obl
nmod
nummod
obl
aux
nmod
Andoni PROPN
izeneko NOUN
mutiko NOUN
bat NUM
baserrian NOUN
bizi ADJ
zen AUX
bere DET
amarekin NOUN
bakar-bakarrik ADV

5

Consecuencias de...

THE NETWORK IN EUROPE

The CLARIN, DARIAH and CLARIAH research infrastructures are active on a national and the European level. CLARIN and DARIAH both have the status of a European Research Infrastructure Consortium (ERIC).

	Polaco	Alemán	Inglés	**Castel**
1. Constituency Parsing		x	x	
2. Coreference Resolution	x			
3. Dependency Parsing	x	x	x	x
4. Distant Reading	x	x	x	x
5. Extraction of Polish terminology	x			
6. Inclusion detection	x			
7. Keyword Extractor	x			
8. Lemmatization		x	x	
9. Machine Translation		x	x	
10. Metadata Processing				
11. Morpho-syntactic tagger	x		x	
12. Morphological Analysis	x	x	x	
13. Named Entity Recognition	x	x	x	x
14. Named Entity Relation Detection				
15. Part-Of-Speech Tagging	x	x	x	
16. Sentiment Analysis	x			
17. Shallow Parsing	x			
18. Spatial expression detection	x			
19. Speech Recognition				
20. Stylometry				
21. TF, IDF, TF-IDF calculation	x			
22. Text Analytics	x	x	x	x
23. Text Enhancement			x	
24. Text Summarization	x			
25. Tokenisation				
26. Topic Modelling				
27. Visualisation of Geographic Data				

Análisis de textos para el español

Más en el Webinario de INTELE: "Análisis de textos para el español" https://github.com/quinnanya/intro-to-nlp-es

Análisis de sentimientos en textos en español y fácilmente adaptable a las lenguas cooficiales

Lección del webinar: https://programminghistorian.org/es/lecciones/analisis-de-sentimientos-r

Parte práctica: https://rstudio.cloud/project/2342606

Distant Reading for European Literary History

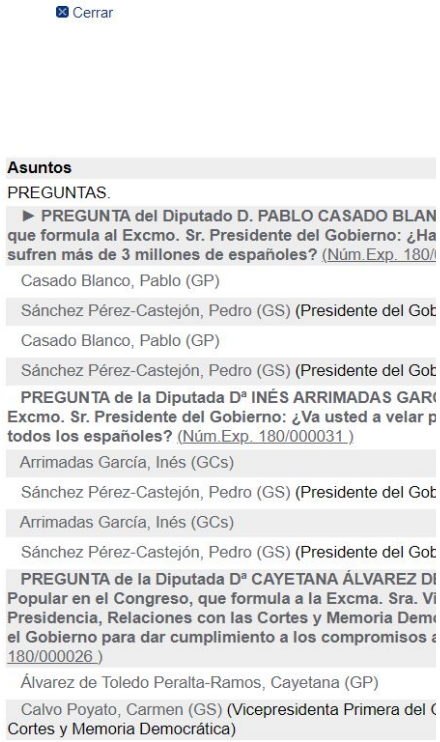
				AUTHORSHIP				LENGTH			TIME SLOT		
Language	Last update	Texts	Words	Male	Female	1-title	3-title	Short	Medium	Long	1840-59	1860-79	1880-99
cze	2021-04-09	100	5621667	88	12	62	6	43	49	8	12	21	39
deu	2021-04-11	100	12738842	67	33	35	9	20	37	43	25	25	25
eng	2021-04-09	100	12227703	49	51	70	10	27	27	46	21	22	31
fra	2021-04-09	100	8712219	66	34	58	10	32	38	30	25	25	25
gsw	2021-06-07	38	2392060	21	17	13	5	11	22	5	0	1	13
hrv	2021-03-22	21	1440018	21	0	4	0	6	12	3	6	12	2
hun	2021-04-09	100	6948590	79	21	71	9	47	31	22	22	21	27
ita	2019-11-21	34	3328244	32	2	19	3	13	10	11	5	12	10

Workshop INTELE: "Distant Reading for European Literary History"

Sesión práctica: github.com/bncolorado/Processing-ELTeC-corpus

Recursos en la nube para la voz

1. Descargar este vídeo del congreso
2. Analizar con un clic en BAS
3. Observar los resultados



1. Mary TTS
2. ASR
3. TextAlign
4. Pipeline without ASR
5. Pho2Syl
6. Chunker
7. AnnotConv
8. G2P
9. OCTRA - online text transcription system.
10. AudioEnhance
11. WebMAUS General
12. Chunk Preparation
13. Coala
14. WebMINNI
15. WebMAUS Basic
16. Anonymizer
17. TextEnhance
18. Formant Analysis
19. Subtitle
20. EMU Magic
21. Voice Activity Detection
22. EMU webApp - online labeling of speech data and more.
23. Pipeline with ASR
24. SpeakDiar

Transcripciones automáticas

Original

(El señor MINISTRO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES (Escrivá Belmonte): Señora López Álvarez,) en primer lugar, permítame que me dirija hacia la pantalla, aunque le dé la espalda, para que se me pueda oír y responda por el señor
Marlaska en términos solidarios, ya que puede ser contestada por el Gobierno. La contestación es clara: no solamente el Gobierno español, yo diría que ningún Gobierno, al menos europeo, en ningún caso fomenta la inmigración irregular, en ningún
caso. Lo que hace es intentar evitar que ocurra. (Aplausos).

IBM

abajo en primer lugar el hermitage hija hacia la pantalla y no puedo a la espalda para que se me pueda oír y respondo por él señor smart laska entrenó solidarios cuyo poder con el gobierno la contestación es clara y no solamente el gobierno español ya que cualquier gobierno almenos europeo en ningún caso fomenta la inmigración irregular ningún caso lo que se intenta evitar que ocurra qué

European Media Lab

señaló que cada año en 1º lugar que permita dirija hacia la pantalla y no pudo dar la espalda para que se me pueda unir y respaldado el señor más hasta que en los aviarios como puede ser colobiano la contestación es clara el no solamente el gobierno español y al igual que el Gobierno al menos europeo y en ningún caso fomenta la inmigración irregular en ningún caso lo que hace es intentar evitar que ocurran

Transcripciones automáticas

Original

(El señor MINISTRO DE INCLUSIÓN, SEGURIDAD SOCIAL Y MIGRACIONES (Escrivá Belmonte): Señora López Álvarez,) en primer lugar, permítame que me dirija hacia la pantalla, aunque le dé la espalda, para que se me pueda oír y responda por el señor
Marlaska en términos solidarios, ya que puede ser contestada por el Gobierno. La contestación es clara: no solamente el Gobierno español, yo diría que ningún Gobierno, al menos europeo, en ningún caso fomenta la inmigración irregular, en ningún
caso. Lo que hace es intentar evitar que ocurra. (Aplausos).

PRAAT textgrid

```
File type = "ooTextFile"

Object class = "TextGrid"



xmin = 0

xmax = 26.302000

tiers? <exists>

(...)

intervals [12]:
        xmin = 3.472000
        xmax = 3.982000
        text = "permita"
    intervals [13]:
        xmin = 3.982000
        xmax = 4.542000
        text = "dirija"
    intervals [14]:
        xmin = 4.542000
        xmax = 4.822000
        text = "hacia"
```

https://youtu.be/7Il-gOShtFA

Transcripciones bilingües euskera-castellano

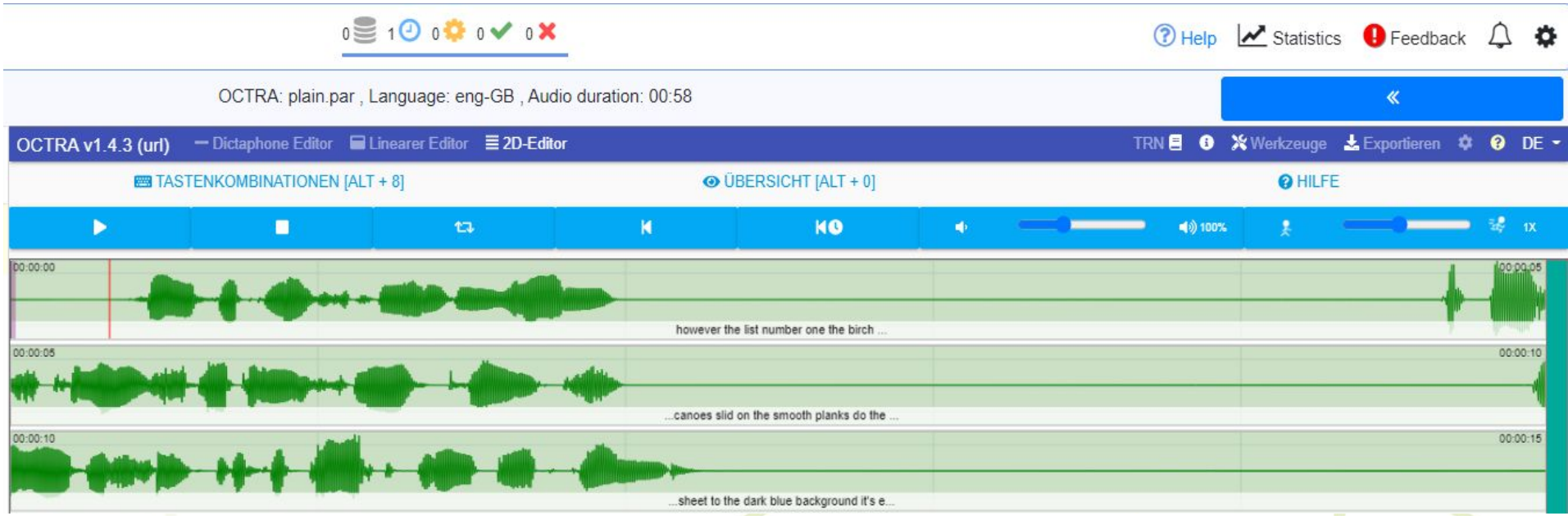
FUENTE:
https://www.legebiltzarra.eus/portal/eu/web/eusko-legebiltzarra/noticias-y-eventos/actos-y-eventos/-/buscador/content/40-aniversario-del-parlamento-vasco-una-mirada-retrospectiva

callGoogleASR: egun on guztioi eta ongi etorri abestia eusko legebiltzarrak euskal herriko unibertsitatearen udako ikastaroen baitan antolatu duen 2000 goiko ikastaro honetara eskerrak eman nahi dizkizuet jardunaldi hauetan parte hartu duzuen guztioi hizlari partehartzaile antolatzaileei ere gehiago covid-19 da gure bizitzak etengabe baldintzatzen dituen une honetan ikastaro hau horren lekuko eusko legebiltzarraren 40. urteurrena atzerako begirada da ikasturte honetarako aukeratutako gaia ezin ziteken besterik izan izan ere aurten 40 (...) izan gara eta legebiltzarrak horretan paper garrantzitsua izan du en estos dos legislaturas el parlamento vasco se ha ido construyendo y consolidando dia a dia del mismo modo que este pueblo nuestro pueblo se ha ido reconstruyendo la trayectoria de la camara ha sido y es fiel reflejo de la evolucion social la presencia de la mujer (...) eta konpromisoz aurre egiteko zuen ekarpenak helburu horretan lagunduko dugula sinetsita berriz ere eskerrak eman nahi dizkizuet guztioi

https://clarin.phonetik.uni-muenchen.de/BASWebServices/data/2021.05.28_00.47.01_234C2EA21FEC82BA69904D1D91E42A41/Eusko-legebiltzarra.txt

OH Portal (CLARIN)

https://oralhistory.eu/oh-portal



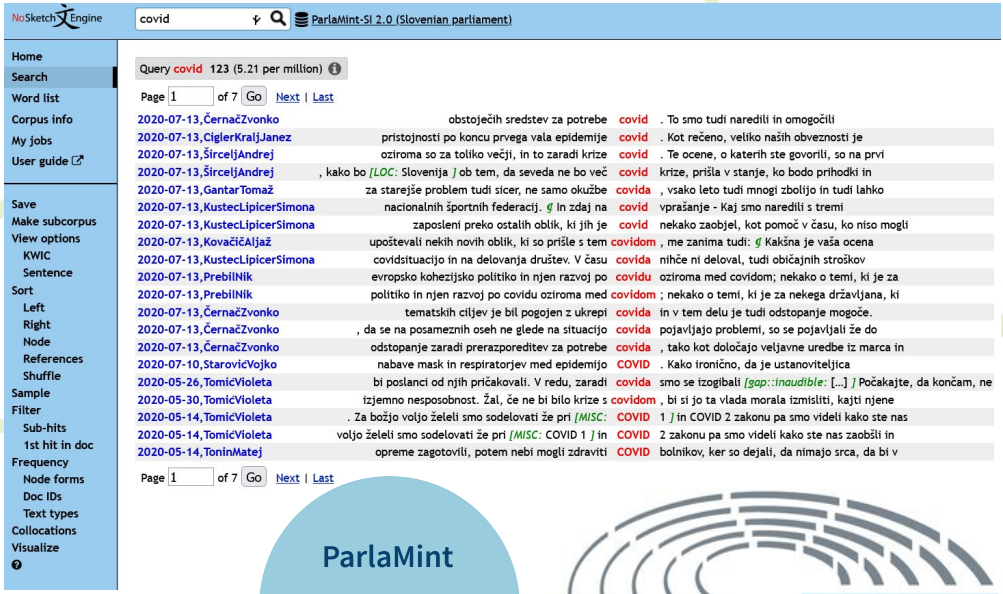
Webinar: https://youtu.be/X6bFGJpMjVQ

https://figshare.com/articles/media/Speech_corpus_-_example_of_raw_audio_HARVARD_list_01_wav/7857770/1

ParlaMint corpus multilingüe (texto escrito)

1. Obtener datos de parlamentos y sus metadatos
2. Convertirlos al esquema de ParlaMint
3. Anotación lingüística (UDpipe y NERC)
4. Hacer corpus disponibles a través de concordantes (noSketch Engine / KonText) y Parlameter

https://www.clarin.si/noske/parlamint.cgi/

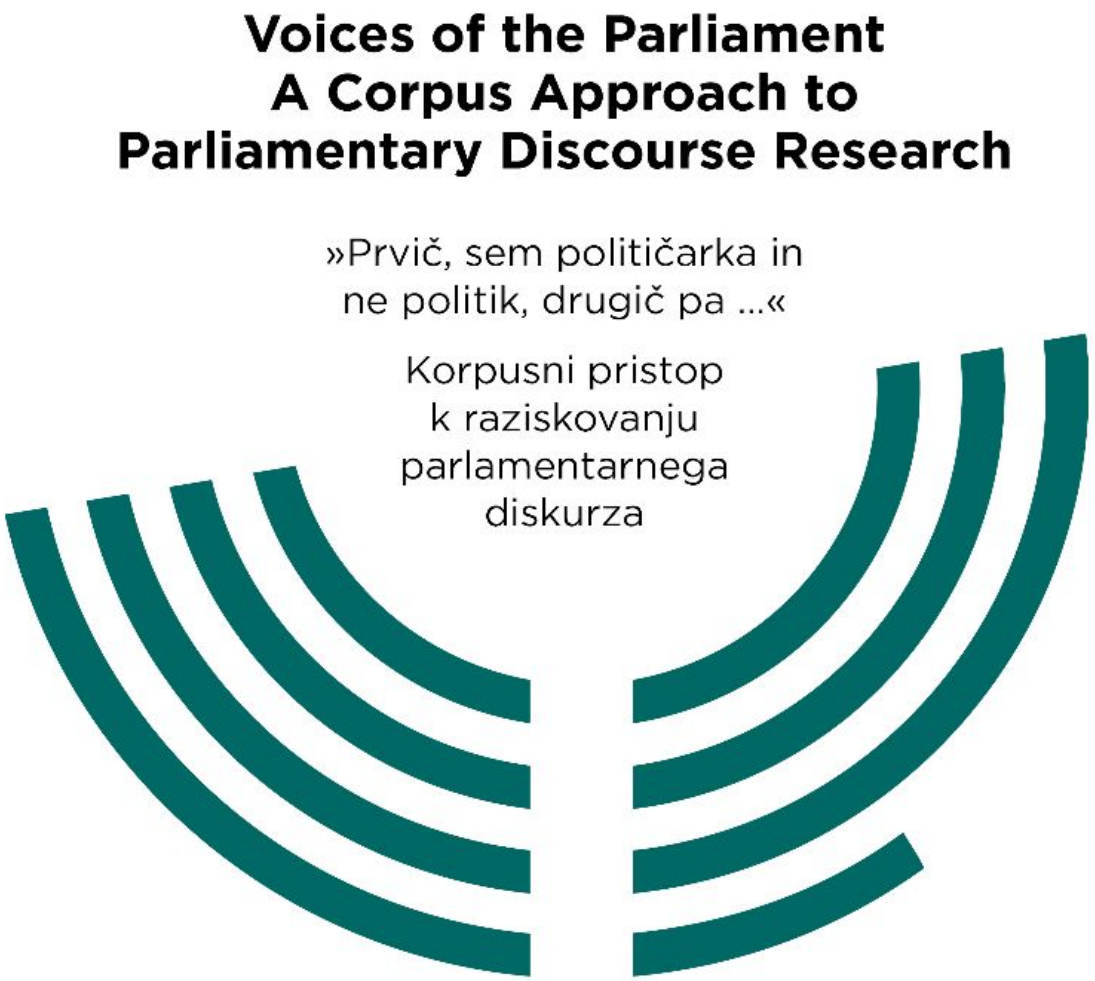


María Calzada es coordinadora de http://blogs.uji.es/ecpc/ y del corpus en castellano de ParlaMint

Preguntas de investigación en el Webinario de INTELE: https://youtu.be/b0oNElZbV9E

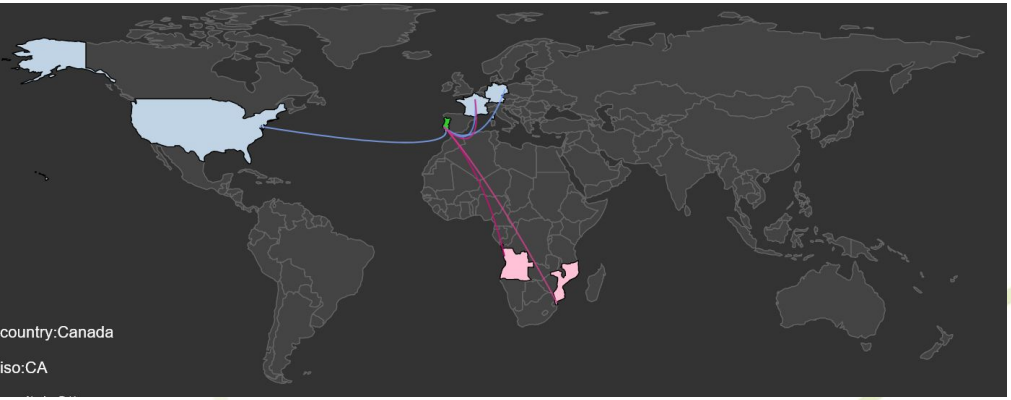
Materiales para el uso práctico del corpus

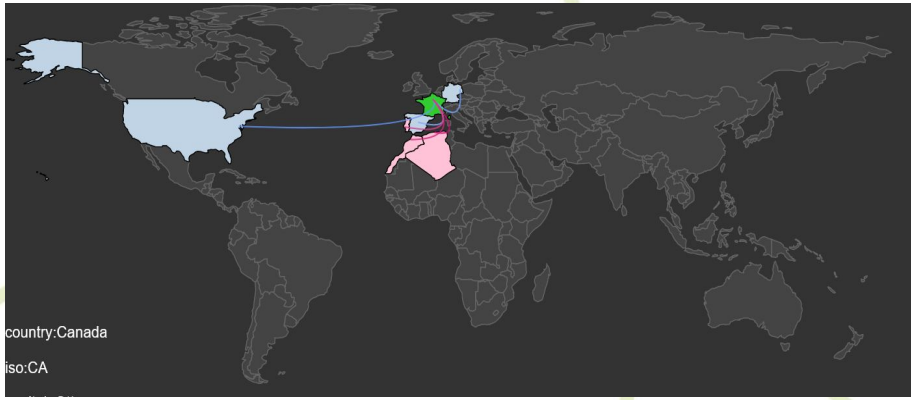
https://sidih.github.io/voices/toc.html

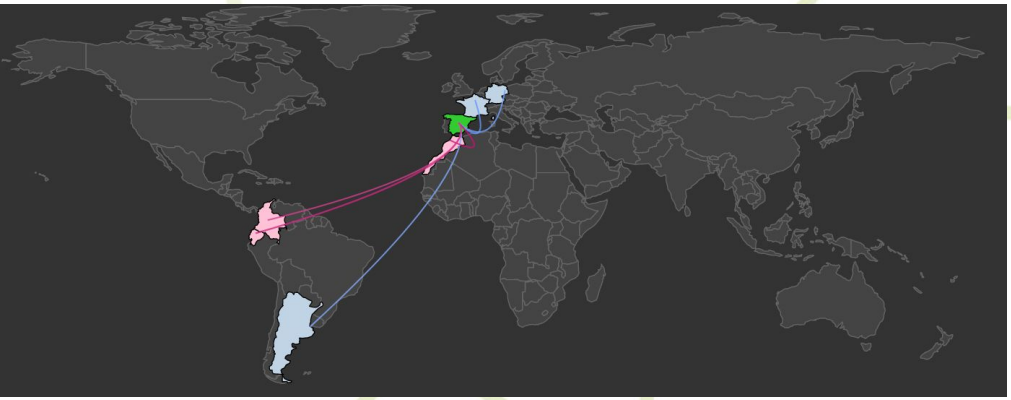


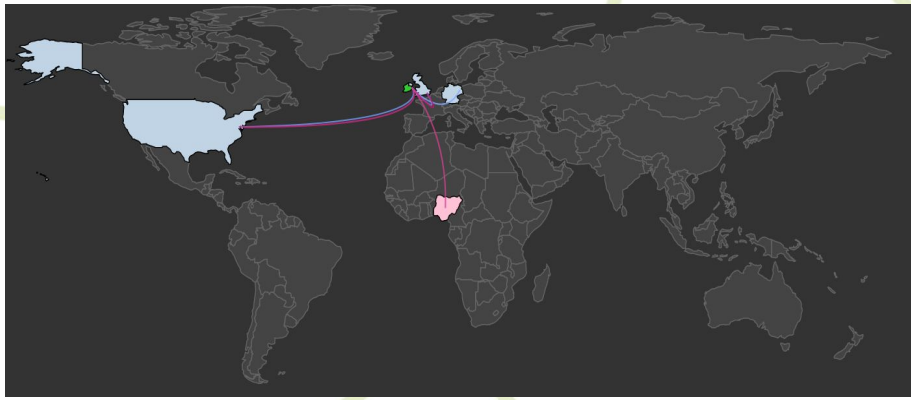
Textual Emigration Analysis

1. Historia, literatura y lingüística computacional.
2. Corpus wikipedia (texto)

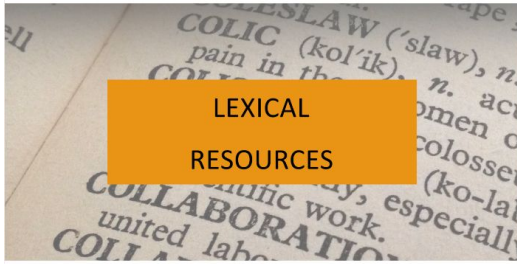






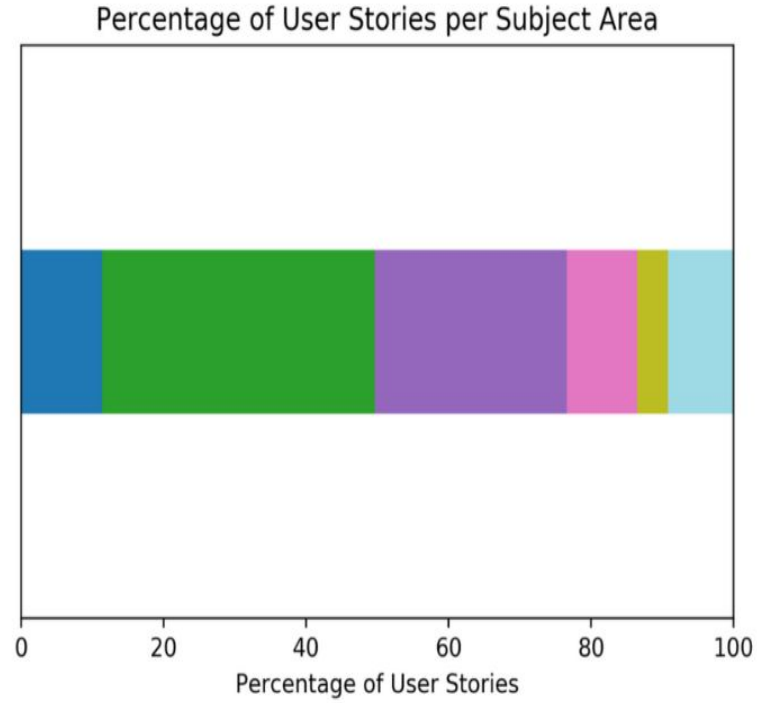


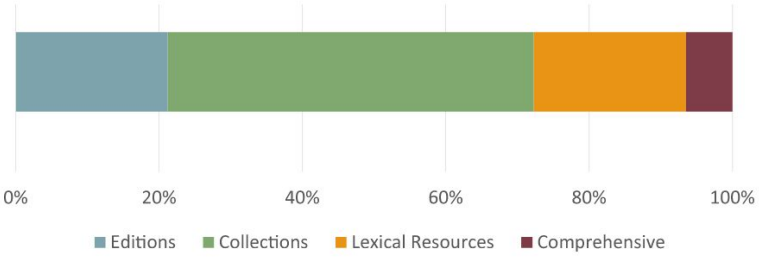
Text+ User Stories (definir necesidades) DARIAH



Text+ User Stories sorted by DFG Subject Areas

Information about the DFG subject areas can be found on the webpages of the DFG.





Distribution bases on 120 user stories on 24 August 2020.

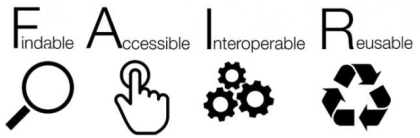
The standarization survival kit DARIAH

- Ayudar en la investigación para utilizar estándares apropiados
- Documentación sobre estándares
- Comunicación entre comunidades investigadoras

Conclusiones de la sesión

- **Ciencia abierta: de impacto y reproducible** (CLARIN)
- Interoperabilidad en (y entre) infraestructuras (EOSC)
  - Construir infraestructura que no se desarrollará en Europa para
    - Lenguas oficiales y cooficiales del estado
  - Evitar la fragmentación: ALL-LT-in-ONE-URL
  - + re-uso de los datos + prosumidores + impacto social
- + Casos de uso y herramientas (sencillas) para investigar

EUROPEAN OPEN SCIENCE CLOUD



CLARIN

Referencias y enlaces de interés

- **Bel, N. Gonzalez-Blanco, E. Iruskieta, M. (2016). CLARIN Centro-K-español. Procesamiento del Lenguaje Natural 57: 151-154. ISSN: 1135-5948.**
- Iruskieta, M. Bel, N. (2017). CLARIN-K Centre Spain: una infraestructura orientada usuario. LINHD-UNED. Escuela de Verano HD.
- **Krauwer, S., & Hinrichs, E. (2014). The CLARIN research infrastructure: resources and tools for e-humanities scholars. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)* (pp. 1525-1531). European Language Resources Association (ELRA).**
- Váradi, T., Wittenburg, P., Krauwer, S., Wynne, M., & Koskenniemi, K. (2008). CLARIN: Common language resources and technology infrastructure. In *6th International Conference on Language Resources and Evaluation (LREC 2008)*.
- Otegi, A. Imaz, O. Díaz de Ilarraza, A. Iruskieta, M. Uria, L. (2017). ANALHITZA: a tool to extract linguistic information from large corpora in Humanities research. Procesamiento del Lenguaje Natural 58, pp. 77-84.
- CLARIN: https://www.clarin.eu/
- DARIAH: https://www.dariah.eu/
- INTELE: http://ixa2.si.ehu.eus/intele/

Eskerrik asko, gracias

- Preguntas??

DH@Madrid Summer School 2021

Herramientas Digitales para las Humanidades Digitales en la e-infraestructura CLARIN

Mikel Iruskieta
Ixa Taldea - HiTZ zentroa
UPV/EHU

www.clarin.eu
www.clarin-es.org
http://ixa2.si.ehu.es/clarink

Creación de un Proyecto de Humanidades Digitales basado en el análisis de textos: Modelado y Procesamiento

UNED
Fundación Uned
LiNHD LABORATORIO DE INNOVACIÓN EN HUMANIDADES DIGITALES
eman ta zabal zazu
Universidad del País Vasco
Euskal Herriko Unibertsitatea
HiTZ
Hizkuntza Teknologiako Zentroa
Basque Center for Language Technology